

Analysing Vote Choice Data

Assignment 1

Jacob Edenhofer*

03 May 2023

Preliminaries

Let us start by importing the necessary packages and data. Please note: I have included the questions in brown to indicate which (sub-)questions a given answer belongs to.

```
# packages
library(tidyverse) # ggplot, dplyr,
library(modelsummary) # summary statistics, regression tables
library(janitor) # data cleaning
library(scales) # add percentages, dates, etc. to plots
library(haven) # importing Stata data
library(here) # relative paths to ensure maximum replicability
library(kableExtra) # tables
library(ggeffects) # obtain and plot predicted values
library(nnet) # multinomial model

# import data
ah21 <- read_dta(paste0(here(), "/Data/ESS2018GER.dta"))
```

Exercise 1

1.1

(a) Get a sense of which variables and how many observations are present in the dataset, and make sure that all the variables are in the correct format: re-scale them when you find it appropriate. You may also want to use function ‘summary()’ to get a sense of how many missing values (if any) are present and if some variables are more problematic than others. (b) Compute the mean of relevant variables and report it in a table. (c) Make sure that all the dichotomous variables are in a “viable” format (0 / 1).

To answer these three questions, I will proceed in two steps:

*jacob.edenhofer@some.ox.ac.uk

- I will transform the two dummy variables, gender (gndr) and (non-)membership of an ethnic minority (blgetmg), into binary (0/1) format to answer part (c).
 - For the purposes of summarising the data, I will not rescale the variables.
- To answer part (b), I will use the modelsummary package's datasummary_skim() function to summarise the data.¹ In doing so, I will treat the ordered² categorical variables (ppltrst, gincdif, freehms, imwbcnt, eduyrs, hinctnta) as numeric variables, whereas I will treat the binary variables and the only non-ordered categorical variable (prtvede1) as factors.

Let us start with the dummy variables then:

```
ah21 <- ah21 %>%
  mutate(gndr1 = ifelse(gndr == 1, 1, 0), # 1 for male, 0 for female
         blgetmg1 = ifelse(blgetmg == 1, 1, 0)) # 1 for ethnic minority, 0 for not
```

Next, I will compute the mean for all non-ordered categorical variables (and other relevant summary statistics) via the following piece of code.³ For emphasis, I have coloured the relevant column in dark blue. It is also worth noting that I have set the histogram argument to FALSE since the resulting histograms are distracting - they are too small to read and not labelled.⁴

```
ah21 %>%
  select(-c(cntry, grep("gndr", names(.)), grep("blgetmg", names(.)), prtvede1)) %>%
  datasummary_skim(output = "kableExtra", histogram = F,
                  title = "Summary statistics for ordered categorical variables") %>%
  kable_styling(latex_options = c("scale_down", "hold_position")) %>%
  column_spec(4, background = "#1F263C", color = "white") %>%
  add_footnote(label = "Source: ESS round 9 (2018)",
              notation = "none")
```

Table 1: Summary statistics for ordered categorical variables

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
Most people can be trusted or you can't be too careful	11	0	5.6	2.2	0.0	6.0	10.0
Government should reduce differences in income levels	6	0	2.1	1.0	1.0	2.0	5.0
Gays and lesbians free to live life as they wish	6	0	1.7	0.8	1.0	2.0	5.0
Immigrants make country worse or better place to live	12	1	5.4	2.2	0.0	5.0	10.0
Years of full-time education completed	25	0	14.8	3.4	5.0	14.0	30.0
Household's total net income, all sources	11	8	6.4	2.7	1.0	7.0	10.0

Source: ESS round 9 (2018)

Apart from the the usual summary statistics (mean, median, standard deviation, minimum and maximum),

¹I prefer this function to the summary() function since it produces roughly the same output, but can be more easily used to produce summary tables via the modelsummary package.

²For ordered categorical variables, higher values correspond to greater/lower support for a given statement; increases in these variables therefore have substantively meaningful interpretations. This is not the case for non-ordered categorical variables, such as prtvede1, where the different values indicate different parties, rather than varying levels of support.

³As can be gleaned from the [ESS codebook](#) for round nine, higher values of ppltrst (first row) indicate greater levels of trust. In contrast, higher values of gincdif indicate less support for redistribution (second row). Similarly, lower values of freehms indicate less support for the LGBT+ people's right to live as they wish. Higher values of imwbcnt indicate greater agreement with the view that immigrants make the country a better place to live (third row).

⁴See the appendix for a version of this table that includes histograms.

the table also shows that there is one variable in particular, `hinctnta`, for which there are significant missing values (116 to be precise).⁵ As a result, some care must be exercised when using the income variable in regressions since item non-response may be correlated with other characteristics that also affect the dependent variable of interest (i.e. confounders).

For the non-ordered categorical variables, I modify the `type` argument, yielding a table with columns indicating the number of observations per category - both in raw terms and as a share of the total number of observations. Hence, I run:

```
ah21 %>%
  select("Gender" = gndr1, "Member of ethnic minority" = blgetmg1,
         "Party voted for in last national election, Germany" = prtvedel1) %>%
  # factorise selected variables, and rename levels for readability
  mutate(Gender = factor(Gender, levels = c("0", "1"),
                        # see code above
                        labels = c("Female", "Male")),
         `Member of ethnic minority` = factor(`Member of ethnic minority`,
                                             levels = c("0", "1"),
                                             labels = c("No", "Yes")),
         `Party voted for in last national election, Germany` =
           factor(`Party voted for in last national election, Germany`,
                 levels = c("1", "2", "3", "4", "5", "6", "7", "8"),
                 labels = c("CDU/CSU", "SPD", "LINKE", "Greens", "FDP",
                           "AfD", "Pirates", "NPD"))) %>%
  datasummary_skim(type = "categorical",
                  output = "kableExtra",
                  title = "Summary statistics for non-ordered categorical variables") %>%
  kable_styling(full_width = T, latex_options = "hold_position") %>%
  column_spec(1, width = "8cm") %>%
  add_footnote(label = "Round 9 of the ESS was run in 2018; 'last national election' therefore refers to",
              notation = "none")
```

Table 2 shows that:

- there are slightly more male respondents than female ones,
- only 3.7% of respondents are members of ethnic minorities, and
- the vote share variables correspond only loosely⁶ to the ones actually obtained by the respective parties in the 2017 national election, the last election before round nine of the ESS in 2018

(d) Represent the distribution of the variable measuring respondents' years of formal education. Label the axes, pick a title, and add a solid, blue line at the median value and a red, dashed line at the mean value.

To answer this question, I run the following code. For the purpose of achieving maximum visual interpretability, I set the `binwidth` argument to two, while the vertical lines are created via the `geom_vline()` functions. The

⁵This number is obtained by running `sum(is.na(ah21$hinctnta))`.

⁶According to the [Bundeswahlleiter](#), the share of party votes ("Zweitstimmenanteile") obtained by the respective parties in 2017 were: 32.9% for the CDU/CSU, 20.5% for the SPD, 12.6% for the AfD, 10.7% for the FDP, 9.2% for the LINKE, and 8.9% for the Greens. The NDP remained well below the five-percent threshold.

Table 2: Summary statistics for non-ordered categorical variables

		N	%
Gender	Female	731	47.7
	Male	800	52.3
Member of ethnic minority	No	1471	96.1
	Yes	57	3.7
Party voted for in last national election, Germany	CDU/CSU	606	39.6
	SPD	400	26.1
	LINKE	107	7.0
	Greens	223	14.6
	FDP	97	6.3
	AfD	96	6.3
	Pirates	1	0.1
	NPD	1	0.1

Round 9 of the ESS was run in 2018; 'last national election' therefore refers to the 2017 German general election.

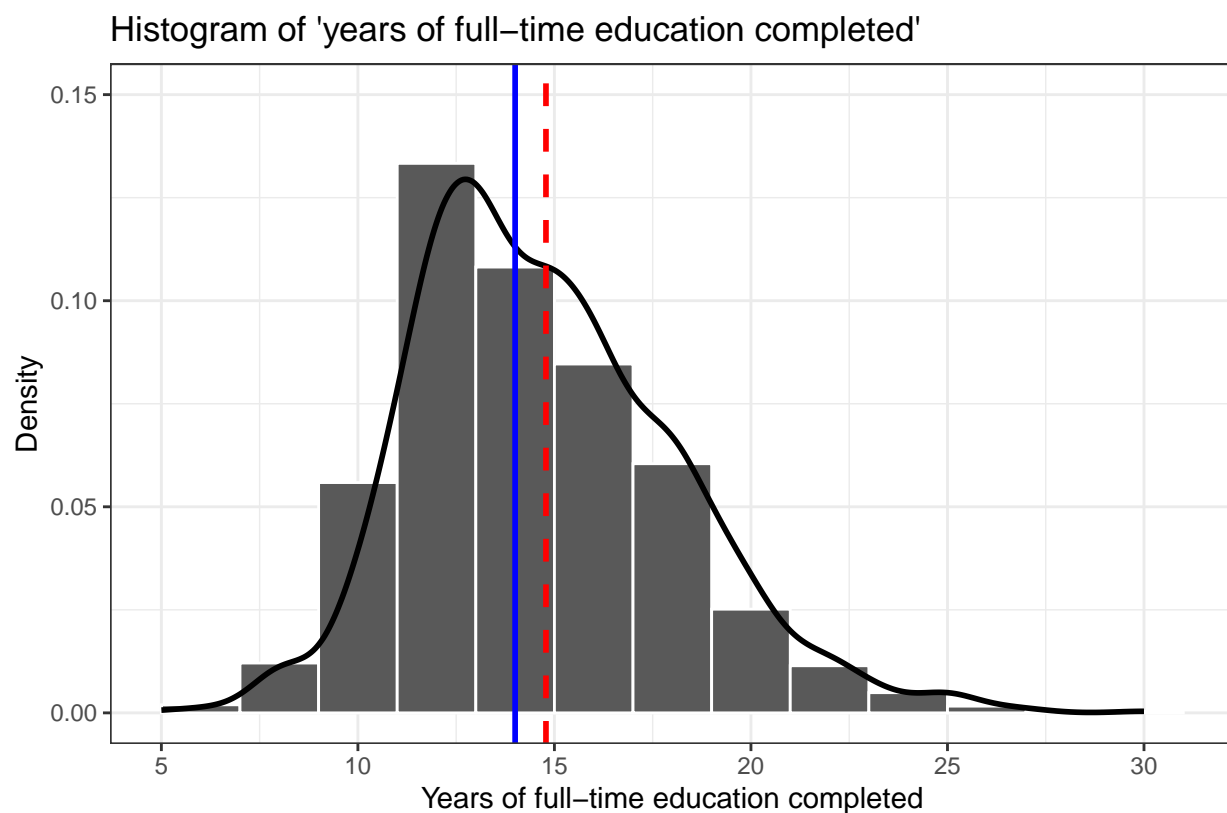
warning messages indicate that there is one missing value for eduyrs, entailing that the histogram is based on 1530 observations.

```
ah21 %>%
```

```
  ggplot(aes(x = eduyrs)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 2, colour = "white") +
  geom_density(linewidth = 1) +
  geom_vline(aes(xintercept = mean(eduyrs, na.rm = T)),
             colour = "red", linetype = "dashed", linewidth = 1) +
  geom_vline(aes(xintercept = median(eduyrs)),
             colour = "blue", linewidth = 1) +
  scale_x_continuous("Years of full-time education completed", breaks = seq(5, 30, 5)) +
  expand_limits(y = 0.15) +
  labs(y = "Density", title = "Histogram of 'years of full-time education completed'",
       caption = "The red dashed line indicates the mean value, while the blue solid line indicates the",
       theme_bw()
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_density()`).
```



What do we learn about the distribution by looking at the mean and the median?

As can be gleaned from the histogram, the mean value (14.8, see table 1) exceeds the median (14, see table 1), implying that the distribution is slightly skewed to the left, though the distribution of `eduyrs` is close to what the theoretical normal distribution with mean 14.8 and a standard deviation of 3.4 (see table 1) would look like.⁷

1.2

I start by creating the dependent variables of interest by, first, filtering out all those parties⁸ that are not contained in table one of the online appendix (Abou-Chadi and Hix 2021) and then dichotomising the remaining four levels:

```
ah21_mod <- ah21 %>%
  filter(!prtved1 %in% c(3, 5, 7, 8)) %>%
  mutate(piketetty_dv = ifelse(prtved1 %in% c(2, 4), 1, 0)) # left equals 1, right equals zero
```

(a) Does the distribution of years of education vary between those voting left and those who do not? For simplicity reasons, group years of education in four possible classes (Hint: use the dummy variable ‘`a la Piketty`’).

⁷The appendix contains a modified version of this figure, which includes the density line of this theoretical normal distribution. The latter is useful for detecting the slight left skew of the actual density line.

⁸Specifically, `ah21_mod` excludes Die LINKE, the NPD, FDP and Pirates.

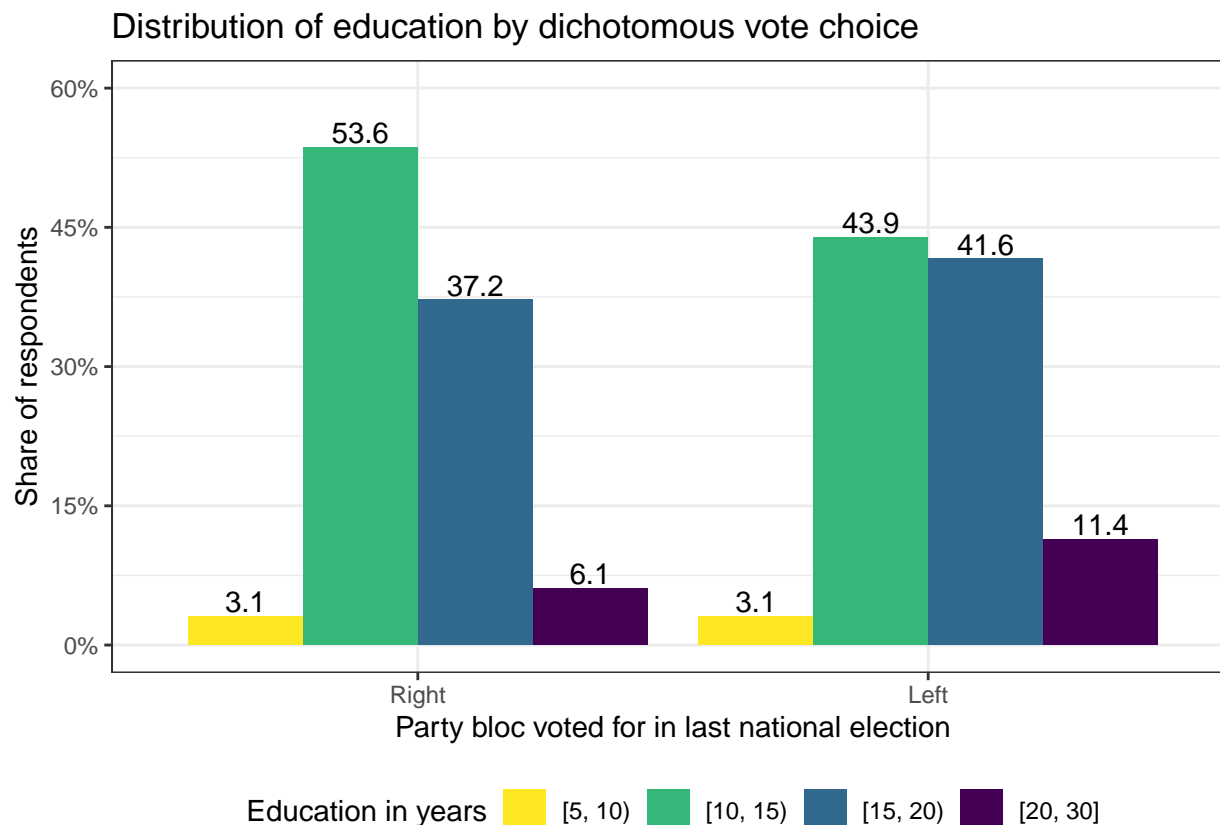
I will, first, create a four-part education variable and, then, plot the distribution of education groups by left-right vote choice. The range of eduyrs is, as table 1 shows, [5,30]. I partition this interval into the following four groups:

- [5,10) corresponds roughly to less than GCSE
- [10,15) corresponds roughly to A-levels and university drop-outs
- [15,20) corresponds roughly to university graduates
- [20,30] corresponds roughly to those with post-graduate qualifications

```
# education variable
ah21_mod <- ah21_mod %>%
  mutate(ed_4part = case_when((eduyrs >= 5 & eduyrs < 10) ~ "[5, 10)",
                              (eduyrs >= 10 & eduyrs < 15) ~ "[10, 15)",
                              (eduyrs >= 15 & eduyrs < 20) ~ "[15, 20)",
                              (eduyrs >= 20 & eduyrs <= 30) ~ "[20, 30]"))

# plot
ah21_mod %>%
  # omit missing observation to avoid cluttered plot
  filter(!is.na(eduyrs)) %>%
  # compute shares in each education group by vote choice
  count(factor(piketty_dv), ed_4part) %>%
  rename("piketty" = `factor(piketty_dv)`) %>% # rename for convenience
  group_by(piketty) %>%
  mutate(share = round(100*(n/sum(n)), 1)) %>%
  # plot this data frame
  ggplot(aes(x = piketty, y = share,
             fill = factor(ed_4part,
                           levels = c("[5, 10)", "[10, 15)",
                                       "[15, 20)", "[20, 30]")))) +
  geom_col(position = "dodge") +
  geom_text(aes(label = share,
                position = position_dodge(width = 0.9),
                vjust = -0.2) +
  scale_fill_viridis_d("Education in years", direction = -1) +
  scale_y_continuous("Share of respondents",
                    breaks = seq(0, 60, 15),
                    labels = label_percent(scale = 1)) +
  scale_x_discrete("Party bloc voted for in last national election",
                  labels = c("0" = "Right",
                             "1" = "Left")) +
  expand_limits(y = 60) +
  labs(x = "Education in years", y = "Share of respondent",
       title = "Distribution of education by dichotomous vote choice") +
  theme_bw() +
```

```
theme(legend.position = "bottom")
```



As the graph above shows, the distribution of education varies in two significant ways between left-bloc voters and right-bloc ones:

- While slightly more than half of all right-bloc voters have between 10 and 15 years of education, this is only the case for 44% of left-bloc voters.
- The proportion of voters with 20 to 30 years of education is almost twice as high among left-bloc voters, compared to their right-bloc counterparts.

(b) Estimate the relationship between voting left and years of education. (Hint: use the dummy variable 'a-la-Piketty') (c) Make sure to write down your estimation equation formally, and to comment the coefficients (d) Why have you chosen this model? How is it different from an OLS model?

To estimate the relationship between voting for the left bloc and years of education, I run a logit model, regressing the dichotomous, Piketty-esque voting variable on eduysrs. I run a logit model, as opposed to an OLS one, because the dependent variable is binary.

By way of justification, let us start with the observation that binary dependent variables imply: $Y \sim \text{Binom}(p)$. We also need to link the parameter, p , to covariates. One idea would be to express p as a linear combination of covariates, $p = x\beta + \epsilon$, where, in the general case, p is n-by-1, x is n-by-p, β is p-by-1 and ϵ is n-by-1. The problem is that this linear probability model fails the range-matching test: p is not bounded between zero and one.

To ensure that for any real-valued set of covariates, p is bounded between zero and one, we use the odds, the probability of an event occurring divided by the probability that it does not occur, $\frac{p}{1-p}$. The odds can take on any non-negative real number. As p increases, the odds increase too. To make sure that the odds are, indeed, non-negative for any covariate value, we write: $\frac{p}{1-p} = \exp(\alpha + \beta X)$.

This link of the model parameter, p , to covariates satisfies our range matching criterion since the exponential function returns positive real numbers for any real-valued input. Solving this expression for p yields: $p = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$. This specification is, however, non-linear in the parameters. To linearise, we use the log - we take the log of the odds (hence, logit), $\log(\frac{p}{1-p}) = \alpha + \beta X$. This is a linear model, not of $E(Y|X)$ (the regression), but of the log of the odds. Thus, logit models are members of the family of generalised linear models⁹, where a non-linear, invertible function, such as $\log(\cdot)$, of the model parameter(s) is expressed as a linear function of the covariates (Gailmard 2014).

With these methodological preliminaries in place, we can write our estimating equation as follows:

$$\log\left(\frac{VoteLeft_i}{1 - VoteLeft_i}\right) = \alpha + \beta YearsEducation_i + \epsilon_i$$

Here, ϵ denotes the error term, while α captures the log odds of voting for the left bloc when an individual has zero years of education. The coefficient of interest is β , the marginal effect of education on the log odds of voting left: the increase in the log odds of voting left for an additional year of completed education.

Since we know that the log odds increase if and only if the probability of voting left increases, we can infer from the sign of $\hat{\beta}$ whether or not an additional year of education increases that probability. We cannot, however, use the coefficient to directly infer the size of the effect.¹⁰ To that end, we use predicted probabilities.

In R, we use the `glm()` function, with the family argument set to binomial, to estimate the above equation. I represent the results both via a conventional regression table and a plot of predicted probabilities obtained via the `ggpredict()` function.

```
# estimate logit
bi_logit <- glm(piketty_dv ~ eduyrs,
               family = binomial(link = "logit"),
               data = ah21_mod)

# regression table
modelsummary(bi_logit,
             estimate = "{estimate}{stars}",
             coef_map = c("eduyrs" = "Years of education"),
             output = "kableExtra",
             title = "Bivariate logit model") %>%
  kable_styling(latex_options = "hold_position")
```

Table 3 shows that an additional year of education is associated¹¹ with a significant¹² increase in the probability

⁹Hence also the `glm()` command in R.

¹⁰Using $p = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$ and applying the quotient as well as chain rules of differentiation, it is straightforward to see that the marginal effect of education on the probability, rather than the log odds, is: $\frac{\partial p}{\partial YearsEducation} = \beta * \frac{\exp(\alpha + \beta YearsEducation)}{(1 + \exp(\alpha + \beta YearsEducation))^2}$.

¹¹I wish to stress that this coefficient is unlikely to capture the causal effect of education on vote voice, given that we do not even control for observed confounders.

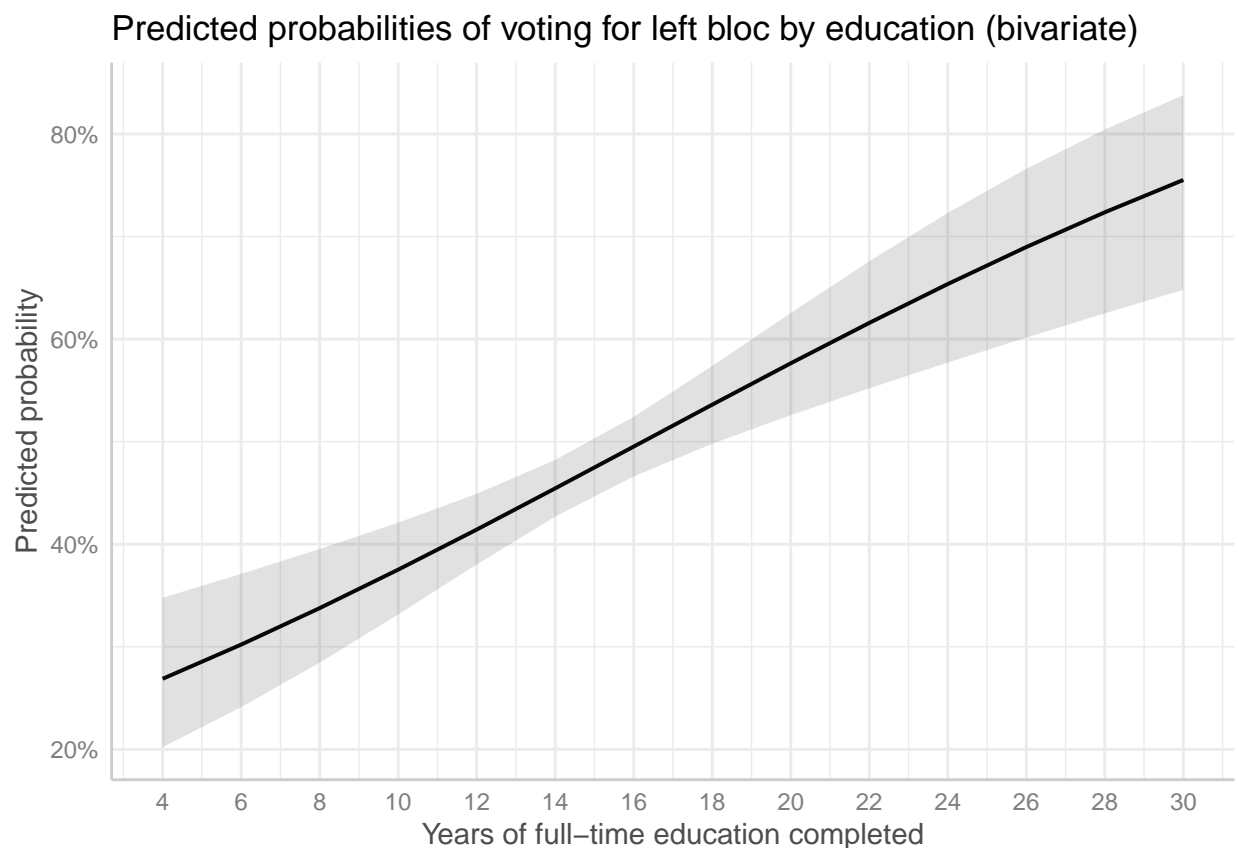
¹²The three stars indicate statistical significance at the 1% level, as explained [here](#).

Table 3: Bivariate logit model

	(1)
Years of education	0.082*** (0.017)
Num.Obs.	1324
AIC	1810.5
BIC	1820.8
Log.Lik.	-903.225
F	23.496
RMSE	0.49

of voting for the left bloc. This conclusion is reinforced by the plot of predicted probabilities below.

```
# plot
plot(ggpredict(bi_logit, terms="eduyrs")) +
  labs(title = "Predicted probabilities of voting for left bloc by education (bivariate)",
       y = "Predicted probability")
```



(e) Run the same model again, but adding the socio-economic covariates and controlling for individuals' level of social capital. Once again, write down the estimation equation formally and do not forget to comment your results after reporting them in a single, tidy table.

In this model, I will add dummies for gender (gndr1) and ethnic minority status (blgetmg1). In addition, trust towards others (ppltrst) will serve as my proxy for social capital, in line with the arguments developed by Putnam (2000). My estimating equation is thus:

$$\log\left(\frac{VoteLeft_i}{1 - VoteLeft_i}\right) = \alpha + \beta_1 YearsEducation_i + \beta_2 Gender_i + \beta_3 EthnicMinority_i + \beta_4 Trust_i + \epsilon_i$$

The estimation of this equation is implemented in R via the following code, which summarises the results in the form of a regression table.

```
# model
multi_logit <- glm(piketty_dv ~ eduyrs + gndr1 + blgetmg1 + ppltrst,
                  family = binomial(link = "logit"),
                  data = ah21_mod)

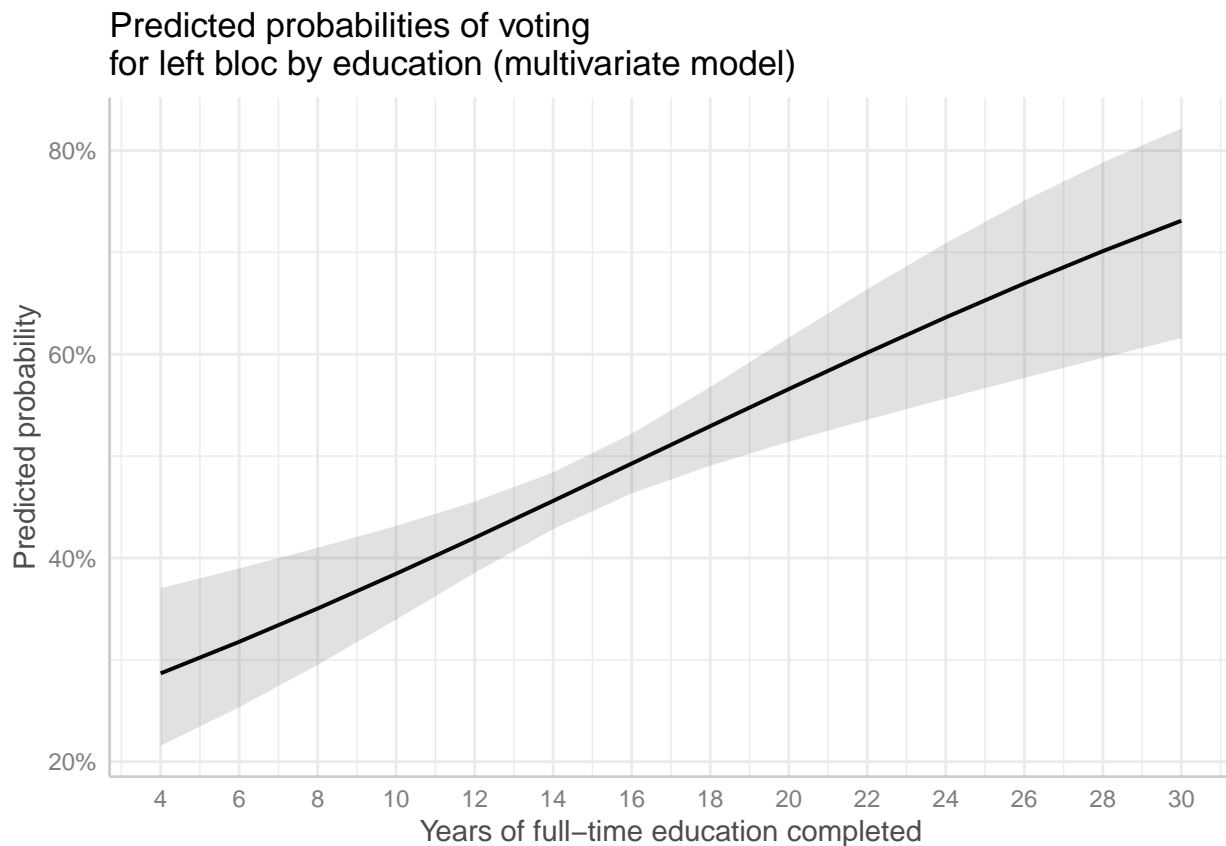
# regression table including both models
modelsummary(list(bi_logit, multi_logit),
             estimate = "{estimate}{stars}",
             coef_map = c("eduyrs" = "Years of education",
                          "gndr1" = "Gender dummy",
                          "blgetmg1" = "Ethnic minority dummy",
                          "ppltrst" = "Trust towards others"),
             output = "kableExtra",
             title = "Multivariate logit model of voting for the left bloc") %>%
kable_styling(latex_options = "hold_position", full_width = T)
```

Table 4: Multivariate logit model of voting for the left bloc

	(1)	(2)
Years of education	0.082*** (0.017)	0.074*** (0.017)
Gender dummy		0.047 (0.113)
Ethnic minority dummy		0.500 (0.314)
Trust towards others		0.071** (0.027)
Num.Obs.	1324	1321
AIC	1810.5	1802.8
BIC	1820.8	1828.7
Log.Lik.	-903.225	-896.382
F	23.496	8.122
RMSE	0.49	0.49

Table 4 demonstrates¹³ that education remains a statistically significant¹⁴ predictor of voting for the left bloc, even after we control for gender, ethnic minority status and social capital. In substantive terms, the coefficient estimate implies that an additional year of education is, ceteris paribus, associated with an increase in the odds¹⁵ of voting for the left bloc by 7.7%. This conclusion is, as above, reinforced by the plot of predicted probabilities below.¹⁶

```
# predicted probability
plot(ggpredict(multi_logit, terms = "eduyrs")) +
  labs(title = "Predicted probabilities of voting\nfor left bloc by education (multivariate model)",
       y = "Predicted probability")
```



1.3

(a) Look back at your results in Exercise 1.2, comment them, and explain how they relate to Abou-Chadi and Hix (2021).

By way of recap, the two key results of the preceding section are:

¹³Furthermore, the coefficient for the 'Gender dummy' captures the difference in voting for the left between men and women, which is not statistically significant. Similarly, the coefficient for the 'Ethnic minority dummy' captures the difference in voting for the left between those belonging to a minority, compared to those who do not. The difference is not statistically significant. 'Trust towards others' is a proxy for social capital; those who are more trusting are, ceteris paribus, also significantly more likely to vote for the left bloc.

¹⁴Significant at the 1% level.

¹⁵This is obtained by computing $100 * (\exp(0.074) - 1)$.

¹⁶As noted above, the marginal effects and, thus, also the predicted probabilities depend on the values of the other covariates. The `ggpredict()` function sets the values of all non-education covariates equal to their mean value.

- The share of those with 20 to 30 years of education among left-bloc voters is almost twice as high as the share among right-bloc voters.
- There is a positive association between years of education and the probability of voting for the left bloc, which is robust to the inclusion of the socio-demographic covariates contained in the dataset.

Broadly speaking, these results are consistent with figure 1 in Abou-Chadi and Hix (2021), though there is an important difference. Using a categorical measure of education, the authors find a non-linear effect of education on left-right vote choice. In our case, the coefficient estimate on the square of years of education is insignificant (see table 11 in the appendix), implying that the effect is positive and linear.

Conceptually, our results are based on what Abou-Chadi and Hix (2021) argue is the “wrong” kind of dependent variable, namely a dichotomous measure of vote choice. Such measures, the authors submit, fail to account for the growing fragmentation of party systems in Western European countries over the last three decades (e.g. Chiaramonte and Emanuele 2019). As a result, analyses resting on binary dependent variables of vote choice do not allow us to say anything about the “mechanisms” of this association. Using a more fine-grained dependent variable, Abou-Chadi and Hix (2021) conclude that the positive association between education and left-wing support is driven by more educated individuals being more likely to vote for left-libertarian, as opposed to mainstream left, parties.

(b) Does education matter? Why? Take into consideration one possible channel through which education affects voting behaviour. Make sure to discuss whether it is a compositional or a contextual effect, that is, whether individuals with certain political attitudes and preferences self-select into/out of education, or whether education shapes and changes political attitudes and preferences.

In (most) advanced industrial democracies, education matters greatly for explaining vote choice in at least two ways. First, as Abou-Chadi and Hix (2021) show, more educated individuals are more likely to support left-libertarian parties, particularly green ones. Secondly, the authors show that “higher educated voters who are in favor of redistribution are more than 30% more likely to support a party of the left than a party of the right. [...] for those “Brahmin” who decide to vote for the left, a key reason for doing so is that they support economic redistribution.” (Abou-Chadi and Hix 2021, 88) That is, when educated individuals hold pro-redistribution preferences, these are a strong motivation for supporting left parties.

Delving deeper into the mechanisms underlying the positive relationship between education and left-wing voting suggests two types of mechanisms are at work. First, education can causally change individuals’ preferences and beliefs. Given that people strategically choose how much education to obtain (self-selection), this mechanism is hard, albeit not impossible, to evaluate empirically. Indeed, several studies exploit compulsory schooling reforms to tease out the causal effect of education on social attitudes and political preferences. Employing such an empirical strategy, Cavaille and Marshall (2019), for instance, find that higher education reduces anti-immigrant sentiment, while Yang (2022) finds that more schooling reduces prejudices against sexual minorities. These studies therefore suggest that education affects individuals’ values, beliefs and preferences in ways that make them more likely to support left-wing, especially left-libertarian, parties.

A second mechanism, however, is at work, namely self-selection into and out of education. That is, we would expect individuals who are ex ante more likely to gain from higher education to strategically choose to obtain more of it, whereas those who are less likely to gain by virtue of, for example, lower motivation, will obtain less education. Self-selection matters for vote choice since (some of) the factors that lead individuals to obtain

more/less education are likely correlated with determinants of vote choice, thus confounding the relationship between education and the latter. Individuals who are more willing to expose themselves to new influences may, for instance, be more likely to attend university, but also view immigration more favourably and, as a result, vote for left-libertarian parties. In that case, it is not education that causes individuals to vote for these parties, but their pre-existing attitudes that explain both university attendance and vote choice.

While I believe there is solid evidence for the causal effect of education on political preferences and behaviour, the effect sizes are small. Moreover, in no Western European country does university attendance significantly exceed 50%, suggesting substantial room for self-selection into and out of education, as, indeed, theories of human capital accumulation would lead us to predict (Becker 2009). Hence, I think that a substantial portion, if not most of it, of the association between education and vote choice is non-causal, i.e. driven by self-selection.

Exercise 2

2.1-2.2

You shall keep using the same dataset from the previous exercise. Test whether there is a positive relationship between voting the Libertarian Left and years of education. Is the aforementioned relationship stronger than that between education and voting for the Mainstream Left? Compare the two specifications in the same tidy table, and explain which criteria you have used in your assessment.

As in exercise 1, I use a logit model to estimate the relationship between the `lib_left_dummy` and `eduyrs`, with the dependent variable being unity when an individual voted for the Greens in the last election and zero otherwise. The result is reported in table 5, along with the results obtained above for the Piketty-like dependent variable.

```
# libertarian left dummy
ah21_mod <- ah21_mod %>%
  mutate(lib_left_dummy = ifelse(prtvedel == 4, 1, 0)) # 1 for Greens, 0 otherwise

# logit model
bi_logit_lib_left <- glm(lib_left_dummy ~ eduyrs,
                        family = binomial(link = "logit"),
                        data = ah21_mod)

# regression table
models1 <- list("Right vs. left bloc" = bi_logit,
               "Left libertarian vs. rest" = bi_logit_lib_left)
modelsummary(models1,
              estimate = "{estimate}{stars}",
              coef_map = c("eduyrs" = "Years of education"),
              output = "kableExtra",
              title = "Comparing the effect of education (bivariate case)") %>%
  kable_styling(latex_options = "hold_position", full_width = T)
```

Table 5: Comparing the effect of education (bivariate case)

	Right vs. left bloc	Left libertarian vs. rest
Years of education	0.082*** (0.017)	0.130*** (0.021)
Num.Obs.	1324	1324
AIC	1810.5	1167.5
BIC	1820.8	1177.9
Log.Lik.	-903.225	-581.760
F	23.496	36.847
RMSE	0.49	0.37

Education, as table 5 bears out, is significantly¹⁷ associated with both voting for the left bloc and for the Greens, the left-libertarian party, though the association is stronger for the latter than for the former. Substantively, an additional year of education is associated with an increase¹⁸ in the odds of voting for the left bloc by approximately 8.5%, whereas the increase is 13.9% for the Greens.

2.3

Restrict the dataset to left-wing voters. Are preferences for income redistribution a good predictor of voting for both the Mainstream Left and the Libertarian Left? Compare your results (again, in the same table, if possible), and discuss them.

To restrict the dataset to left-wing voters, I consider only voters who voted for the SPD (`prtvede1 == 2`) or the Greens (`prtvede1 == 4`) since these are the only left-wing parties listed in table 1 of the online appendix (Abou-Chadi and Hix 2021). Then, I regress `m_left_dummy` and `left_lib_dummy` respectively on `gincdif`, and estimate these specifications via a logit model. In these two models, respondents' (dis)agreement with the statement that 'government should reduce income differences' serves as our proxy for attitudes towards redistribution (`gincdif`).

```
# prune data to left-wing voters only
ah21_mod_left <- ah21_mod %>%
  filter(prtvede1 %in% c(2, 4)) %>% # Greens and SPD
  mutate(left_lib_dummy = ifelse(prtvede1 == 4, 1, 0), # Greens
         m_left_dummy = ifelse(prtvede1 == 2, 1, 0)) # SPD

# models
bi_logit_re_ml <- glm(m_left_dummy ~ gincdif,
                     family = binomial(link = "logit"),
                     data = ah21_mod_left)
bi_logit_re_ll <- glm(left_lib_dummy ~ gincdif,
                     family = binomial(link = "logit"),
                     data = ah21_mod_left)
```

¹⁷Significant at the 1% level

¹⁸As explained above, these values are obtained by computing $100 \cdot (\exp(0.082) - 1)$ and $100 \cdot (\exp(0.13) - 1)$.

```
# modelsummary
models2 <- list("Mainstream left (SPD)" = bi_logit_re_ml,
               "Left libertarian (Greens)" = bi_logit_re_ll)

# table
modelsummary(models2,
              estimate = "{estimate}{stars}",
              coef_map = c("gincdif" = "Gov't should reduce income differences"),
              title = "Attitudes towards redistribution as predictor of vote choice among left-wing voters",
              output = "kableExtra") %>%
  kable_styling(latex_options = "hold_position", full_width = T) %>%
  add_footnote(label = "See footnote 3 for interpreting the explanatory variable.",
              notation = "none")
```

Table 6: Attitudes towards redistribution as predictor of vote choice among left-wing voters (bivariate)

	Mainstream left (SPD)	Left libertarian (Greens)
Gov't should reduce income differences	-0.022 (0.094)	0.022 (0.094)
Num.Obs.	623	623
AIC	816.6	816.6
BIC	825.5	825.5
Log.Lik.	-406.311	-406.311
F	0.053	0.053
RMSE	0.48	0.48

See footnote 3 for interpreting the explanatory variable.

Table 6 shows that - when restricting the data only to left-wing voters - attitudes towards redistribution are not a statistically significant predictor of vote choice. This does not imply, it is worth noting, that redistributive preferences do not affect the probability of supporting left-wing parties in general (i.e. among all voters).¹⁹ Instead, this result is best interpreted as showing that those with relatively strong preferences for redistribution 'select' into the left electorate. As a result, redistributive preferences do not vary all that much among left-wing voters, explaining our 'null' finding.

2.4

What happens when you add attitudes towards immigrants and sexual minorities as covariates? (a) Report the estimation equation formally, (b) Comment the coefficients, reporting the two model specifications in the same, tidy table, (c) What do you learn from this? Are preferences for redistribution always a good predictor of left voting?

I will run two logit models - one with `m_left_dummy` as the dependent variable, and the other with `left_lib_dummy`. Letting p denote the respective probabilities, we can write the estimating equation as follows:

¹⁹Table 12 in the appendix demonstrates this point explicitly. In models (2) and (3), redistributive attitudes are strongly predictive of voting for the left bloc.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 \text{YearsEducation}_i + \beta_2 \text{Gender}_i + \beta_3 \text{ImmigrantSentiment}_i + \beta_4 \text{LGBT}_i + \epsilon_i$$

As above, our main coefficient of interest is β_1 , which represents the marginal effect of an additional year on education on the log odds of voting for the mainstream left or left-libertarian parties, holding all other variables constant. Formally, we can write this as:

$$\frac{\partial \log\left(\frac{p_i}{1-p_i}\right)}{\partial \text{YearsEducation}} = \beta_1$$

More intuitively, we can interpret $100 * (\exp(\beta_1) - 1)$ as the percent change in the odds for an additional year of education. The interpretation of all other coefficients is analogous. They all represent partial derivatives.

In R, I estimate this equation by running:

```
# models
bi_logit_re_ml_m <- glm(m_left_dummy ~ gincdif + imwbcnt + freehms,
                        family = binomial(link = "logit"),
                        data = ah21_mod_left)
bi_logit_re_ll_m <- glm(left_lib_dummy ~ gincdif + gincdif + imwbcnt + freehms,
                        family = binomial(link = "logit"),
                        data = ah21_mod_left)

# modelsummary
models3 <- list("Mainstream left (SPD)" = bi_logit_re_ml_m,
               "Left libertarian (Greens)" = bi_logit_re_ll_m)

# table
modelsummary(models3,
              estimate = "{estimate}{stars}",
              coef_map = c("gincdif" = "Gov't should reduce income differences",
                           "imwbcnt" = "Immigrants make country worse/better place to live",
                           "freehms" = "LBGT+ people should be free to live as they wish"),
              title = "Attitudes towards redistribution as predictor of vote choice among left-wing voters",
              output = "kableExtra") %>%
kable_styling(latex_options = "hold_position", full_width = T)
```

As before, the coefficient estimate on redistributive preferences is not statistically significant, suggesting that, among left-wing voters, redistributive preferences are not a strong predictor of which type of left-wing party they vote for.²⁰ This suggests that individuals with strong redistributive preferences select into the left party bloc, but other variables then determine which left party they vote for.

²⁰See table 12 in the appendix, which shows that, among all voters, redistributive attitudes do significantly predict left-bloc vote choice, even after controlling for socio-demographic and attitudinal variables.

Table 7: Attitudes towards redistribution as predictor of vote choice among left-wing voters (multivariate)

	Mainstream left (SPD)	Left libertarian (Greens)
Gov't should reduce income differences	-0.145 (0.101)	0.145 (0.101)
Immigrants make country worse/better place to live	-0.219*** (0.047)	0.219*** (0.047)
LGBT+ people should be free to live as they wish	0.650*** (0.155)	-0.650*** (0.155)
Num.Obs.	622	622
AIC	762.9	762.9
BIC	780.6	780.6
Log.Lik.	-377.439	-377.439
F	15.213	15.213
RMSE	0.46	0.46

2.5

(a) Go back to the main dataset: you want to explore the relationship between preferences for redistribution and voting behavior in general: (a) What model would you want to run? And why? (b) Run the appropriate model (with and without the aforementioned attitudinal covariates) and report your baseline category of choice. What is the point of having a baseline category? Why do MNL models have it while logit models do not?

Exploring the relationship between redistributive preferences and voting behaviour in general is best done by means of a multinomial model, which accounts for our dependent variable being an unordered categorical variable consisting of the different parties. Multinomial models therefore address one shortcoming of logit models. With an unordered categorical variable (with more than two levels), we can only estimate a logit model by collapsing this variable. This is potentially problematic since our reference group (the “zero” group) may consist of highly heterogeneous entities, rendering a substantively meaningful interpretation difficult, if not impossible.

To see this, suppose we regressed a dummy for SPD voting on redistributive preferences and found a positive coefficient. The interpretation would then be that a stronger preference for redistribution increases the probability of voting for the SPD, relative to not doing so. But “not doing so” includes voting for the CDU/CSU, FDP, NPD, LINKE or AfD - a highly disparate group of parties.

Hence, I will run two multinomial models without and with attitudinal covariates, where my reference category is voting for the SPD - the mainstream left party. Given that multinomial models have multi-valued categorical variable as their dependent variables, they require a reference category, unlike logit models, where the reference group is the “zero” group. I choose this reference category because the debate between Piketty, on the one hand, and Abou-Chadi and Hix, on the other, mainly concerns the fortunes of this party family.

In R, I estimate the models via the following code. Given the relatively complex output of the `multinom()` function, I only represent the coefficient estimates for `gincdif`, and do so for each model separately.²¹

²¹Unfortunately, I have not been able to suppress the messages printed by `multinom()`.

```

# set SPD reference category
ah21$prtvede11 <- relevel(factor(ah21$prtvede1), ref = '2')

# estimate models
multinom1 <- multinom(prtvede11 ~ gincdif,
                      data = ah21)

```

weights: 24 (14 variable)

initial value 3177.386676 iter 10 value 2501.351614 iter 20 value 2313.757002 iter 30 value 2311.405825 iter 40 value 2311.127502 iter 50 value 2310.830344 final value 2310.821418 converged

```

multinom2 <- multinom(prtvede11 ~ gincdif + freehms + imwbcnt,
                      data = ah21)

```

weights: 40 (28 variable)

initial value 3146.195053 iter 10 value 2397.929738 iter 20 value 2193.150798 iter 30 value 2141.544175 iter 40 value 2133.250099 iter 50 value 2132.805092 iter 60 value 2132.691336 iter 70 value 2132.676187 iter 80 value 2132.668315 iter 90 value 2132.665554 iter 100 value 2132.659997 final value 2132.659997 stopped after 100 iterations

```

models3 <- list(multinom1, multinom2)

# loop for printing out tables
for(i in c(1:2)){
  dd <- modelsummary(models3[[i]],
                    estimate = "{estimate}{stars}",
                    coef_map = c("gincdif" = "Attitudes towards redistribution",
                                "feehms" = "LGBT+ people should be free to live as they wish",
                                "imwbcnt" = "Immigrants make the country a worse/better place to live"),
                    output = "data.frame") %>%
  filter(grepl("Attitudes towards redistribution", term),
         grepl("estimate", statistic))
  # add rownames, derived from ESS codebook (CDU/CSU = 1, LINKE = 3, Greens = 4, FDP = 5, AfD = 6, Pirates = 7, NPD = 8)
  rownames(dd) <- c("CDU/CSU", "LINKE", "Greens", "FDP", "AfD", "Pirates", "NPD")
  dd <- dd %>%
  select(-1) %>%
  kbl(booktabs = T,
      caption = paste0("Attitudes towards redistribution based on model ", i)) %>%
  kable_styling(full_width = T)
  print(dd)
}

```

Table 8: Attitudes towards redistribution based on model 1

	term	statistic	(1)
CDU/CSU	Attitudes towards redistribution	estimate	-0.498***
LINKE	Attitudes towards redistribution	estimate	-10.856***
Greens	Attitudes towards redistribution	estimate	0.021
FDP	Attitudes towards redistribution	estimate	0.045
AfD	Attitudes towards redistribution	estimate	0.248*
Pirates	Attitudes towards redistribution	estimate	0.411***
NPD	Attitudes towards redistribution	estimate	0.533***

Table 9: Attitudes towards redistribution based on model 2

	term	statistic	(1)
CDU/CSU	Attitudes towards redistribution	estimate	-0.484**
LINKE	Attitudes towards redistribution	estimate	-12.947***
Greens	Attitudes towards redistribution	estimate	0.004
FDP	Attitudes towards redistribution	estimate	0.110
AfD	Attitudes towards redistribution	estimate	0.279*
Pirates	Attitudes towards redistribution	estimate	0.398***
NPD	Attitudes towards redistribution	estimate	0.551***

Finally, let us briefly interpret the coefficient estimates. Since I have not rescaled the redistributive preference variable an increase in the latter, recall, amounts to less support for redistribution. So, the statistically significant coefficient estimates on CDU/CSU and LINKE mean that, as individuals become less supportive of redistribution, they are less likely to vote for these two parties, relative to voting for the SPD. The effect is particularly pronounced for the LINKE, suggesting that people with moderate redistributive preferences are, *ceteris paribus*, much more likely to vote for the SPD than for the LINKE. Positive coefficient estimates, by contrast, mean that, as individuals become less supportive of redistribution, they are more likely to vote for the AfD, the NPD, or the Pirates, relative to voting for the SPD. When comparing the SPD to the FDP and Greens, there is no significant effect of redistributive preferences on vote choice.



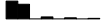



Appendix

Exercise 1

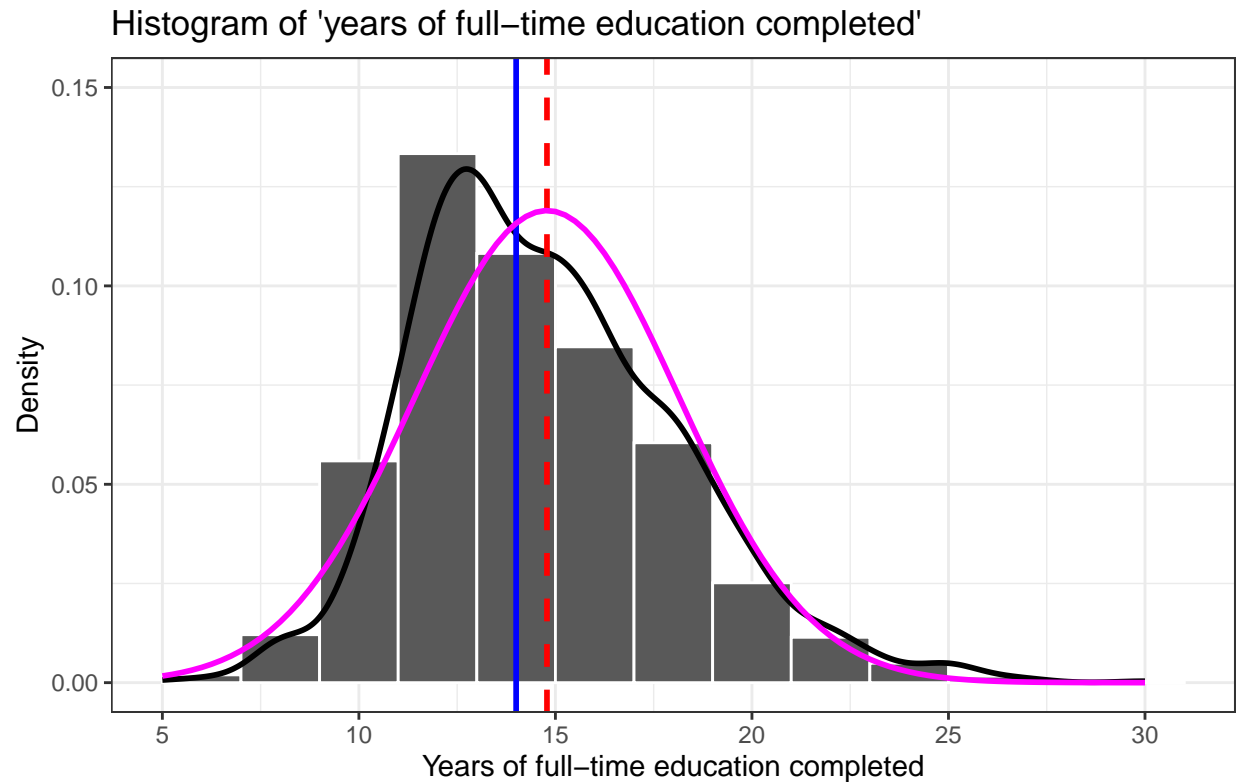
1.1

```
ah21 %>%
  select(-c(cntry, gndr, gndr1, blgetmg, blgetmg1, prtvedel)) %>%
  datasummaryskim(output = "kableExtra",
                  title = "Summary statistics for ordered categorical variables") %>%
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

Table 10: Summary statistics for ordered categorical variables

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
Most people can be trusted or you can't be too careful	11	0	5.6	2.2	0.0	6.0	10.0	
Government should reduce differences in income levels	6	0	2.1	1.0	1.0	2.0	5.0	
Gays and lesbians free to live life as they wish	6	0	1.7	0.8	1.0	2.0	5.0	
Immigrants make country worse or better place to live	12	1	5.4	2.2	0.0	5.0	10.0	
Years of full-time education completed	25	0	14.8	3.4	5.0	14.0	30.0	
Household's total net income, all sources	11	8	6.4	2.7	1.0	7.0	10.0	

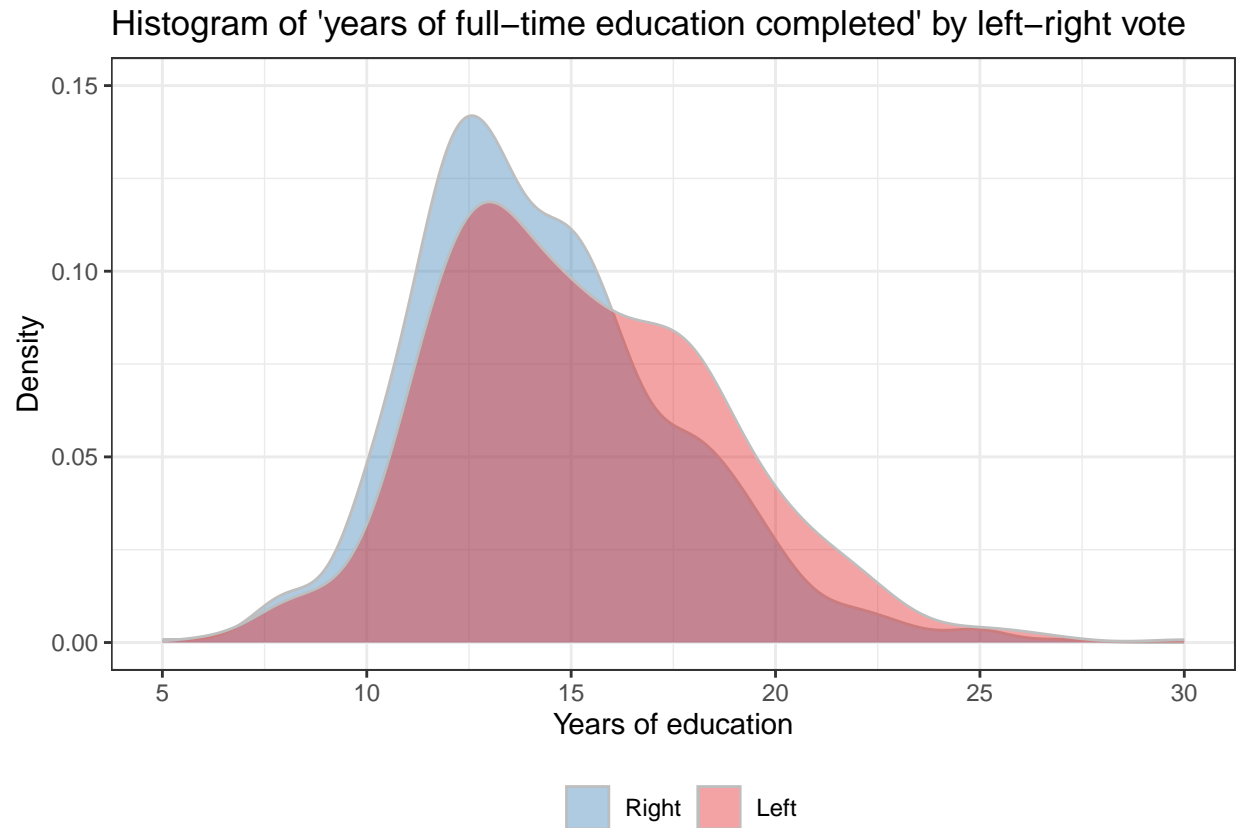
```
ah21 %>%
  ggplot(aes(x = eduyrs)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 2, colour = "white") +
  geom_density(size = 1) +
  geom_vline(aes(xintercept = mean(eduyrs, na.rm = T)),
            colour = "red", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median(eduyrs)),
            colour = "blue", size = 1) +
  stat_function(fun = dnorm,
              args = list(mean = mean(ah21$eduyrs, na.rm = T),
                          sd = sd(ah21$eduyrs, na.rm = T)),
              colour = "magenta", size = 1) +
  scale_x_continuous("Years of full-time education completed", breaks = seq(5, 30, 5)) +
  expand_limits(y = 0.15) +
  labs(y = "Density", title = "Histogram of 'years of full-time education completed'",
       caption = "The red dashed line indicates the mean value, while the blue solid\nline indicates the th",
       theme_bw()
```



The red dashed line indicates the mean value, while the blue solid line indicates the median. The magenta line indicates the theoretical normal distribution.

1.2

```
ah21_mod %>%
  ggplot(aes(x = edu yrs, fill = factor(piketty_dv))) +
  geom_density(alpha = 0.4, colour = "gray") +
  scale_fill_brewer("", palette = "Set1", direction = -1,
    labels = c("0" = "Right",
               "1" = "Left")) +
  scale_x_continuous("Years of education",
    breaks = seq(5, 30, 5)) +
  scale_y_continuous("Density",
    breaks = seq(0, 0.15, 0.05)) +
  expand_limits(y = 0.15) +
  labs(title = "Histogram of 'years of full-time education completed' by left-right vote") +
  theme_bw() +
  theme(legend.position = "bottom")
```



1.3

```
# estimate logit
bi_logit_nl <- glm(piketty_dv ~ eduyrs + I(eduyrs^2),
  family = binomial(link = "logit"),
  data = ah21_mod)

# regression table
modelsummary(bi_logit_nl,
  estimate = "{estimate}{stars}",
  coef_map = c("eduyrs" = "Years of education",
    "I(eduyrs^2)" = "Years of education squared"),
  output = "kableExtra",
  title = "Logit model with non-linear education effect") %>%
  kable_styling(latex_options = "hold_position")
```

Exercise 2

2.3

```
# models
multi_logit <- glm(piketty_dv ~ eduyrs + gndr1 + blgetmg1 + ppltrst + gincdif,
```

Table 11: Logit model with non-linear education effect

	(1)
Years of education	0.020 (0.112)
Years of education squared	0.002 (0.004)
Num.Obs.	1324
AIC	1812.1
BIC	1827.7
Log.Lik.	-903.067
F	11.790
RMSE	0.49

```

family = binomial(link = "logit"),
data = ah21_mod)

multi_logit1 <- glm(piketty_dv ~ eduyrs + gndr1 + blgetmg1 + ppltrst +
  gincdif + freehms + imwbcnt,
  family = binomial(link = "logit"),
  data = ah21_mod)

# regression table including both models
modelsummary(list(bi_logit, multi_logit, multi_logit1),
  estimate = "{estimate}{stars}",
  coef_map = c("eduyrs" = "Years of education",
    "gndr1" = "Gender dummy",
    "blgetmg1" = "Ethnic minority dummy",
    "ppltrst" = "Trust towards others",
    "gincdif" = "Attitudes towards redistribution",
    "feehms" = "LGBT+ people should be free to live as they wish",
    "imwbcnt" = "Immigrants make the country a worse/better place to live"),
  output = "kableExtra",
  title = "Comparing bi- and multivariate logit models of voting for the left bloc") %>%
kable_styling(latex_options = "hold_position", full_width = T) %>%
add_footnote(label = "See footnote 3 for interpreting the attitudinal covariates.", notation = "none")

```

Table 12: Comparing bi- and multivariate logit models of voting for the left bloc

	(1)	(2)	(3)
Years of education	0.082*** (0.017)	0.079*** (0.018)	0.044* (0.019)
Gender dummy		0.080 (0.115)	0.162 (0.120)
Ethnic minority dummy		0.518 (0.321)	0.707* (0.347)
Trust towards others		0.074** (0.027)	0.008 (0.029)
Attitudes towards redistribution		−0.401*** (0.060)	−0.355*** (0.062)
Immigrants make the country a worse/better place to live			0.201*** (0.030)
Num.Obs.	1324	1318	1305
AIC	1810.5	1753.0	1663.8
BIC	1820.8	1784.1	1705.2
Log.Lik.	−903.225	−870.515	−823.881
F	23.496	14.805	18.998
RMSE	0.49	0.48	0.47

See footnote 3 for interpreting the attitudinal covariates.

References

- Abou-Chadi, Tarik, and Simon Hix. 2021. “Brahmin Left Versus Merchant Right? Education, Class, Multiparty Competition, and Redistribution in Western Europe.” *The British Journal of Sociology* 72 (1): 79–92. <https://doi.org/10.1111/1468-4446.12834>.
- Becker, Gary S. 2009. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Chicago: University of Chicago Press.
- Cavaille, Charlotte, and John Marshall. 2019. “Education and Anti-Immigration Attitudes: Evidence from Compulsory Schooling Reforms Across Western Europe.” *American Political Science Review* 113 (1): 254–63.
- Chiaramonte, Alessandro, and Vincenzo Emanuele. 2019. “Towards Turbulent Times: Measuring and Explaining Party System (de-) Institutionalization in Western Europe (1945–2015).” *Italian Political Science Review / Rivista Italiana Di Scienza Politica* 49 (1): 1–23.
- Gailmard, Sean. 2014. *Statistical Modeling and Inference for Social Science*. Cambridge: Cambridge University Press.
- Putnam, Robert D. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon; schuster.
- Yang, Songtao. 2022. “More Education, Less Prejudice Against Sexual Minorities? Evidence from Compulsory Schooling Reforms.” *Applied Economics Letters* 29 (19): 1840–46.