GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Dr. Christian Czymara

FORSCHUNGSPRAKTIKUM I UND II: LÄNGSSCHNITTDATENANALYSE IN R

Linear and non-linear probability models
session vii

AGENDA

- So far we discussed continuous depended variables
- Today we will focus on binary ~~and multi-categorical~~ outcomes
- … And the Maximum Likelihood estimator

DUMMY VARIABLES

WHAT ARE DUMMY VARIABLES

- Binary variables (values 0 and 1)
- Identifying categories
- Dummy variables may be used (kind of) like continuous variables

CODING OF DUMMY VARIABLES

- Mean equals proportion of observations belonging to category
- Or probability of observing category

→both have a continuous character

```
> prop.table(table(data$female))

        0         1
0.6227273 0.3772727
> mean(data$female)
[1] 0.3772727
```

→~38 percent female, so ~62 percent not female

DUMMY VARIABLES AS OUTCOMES

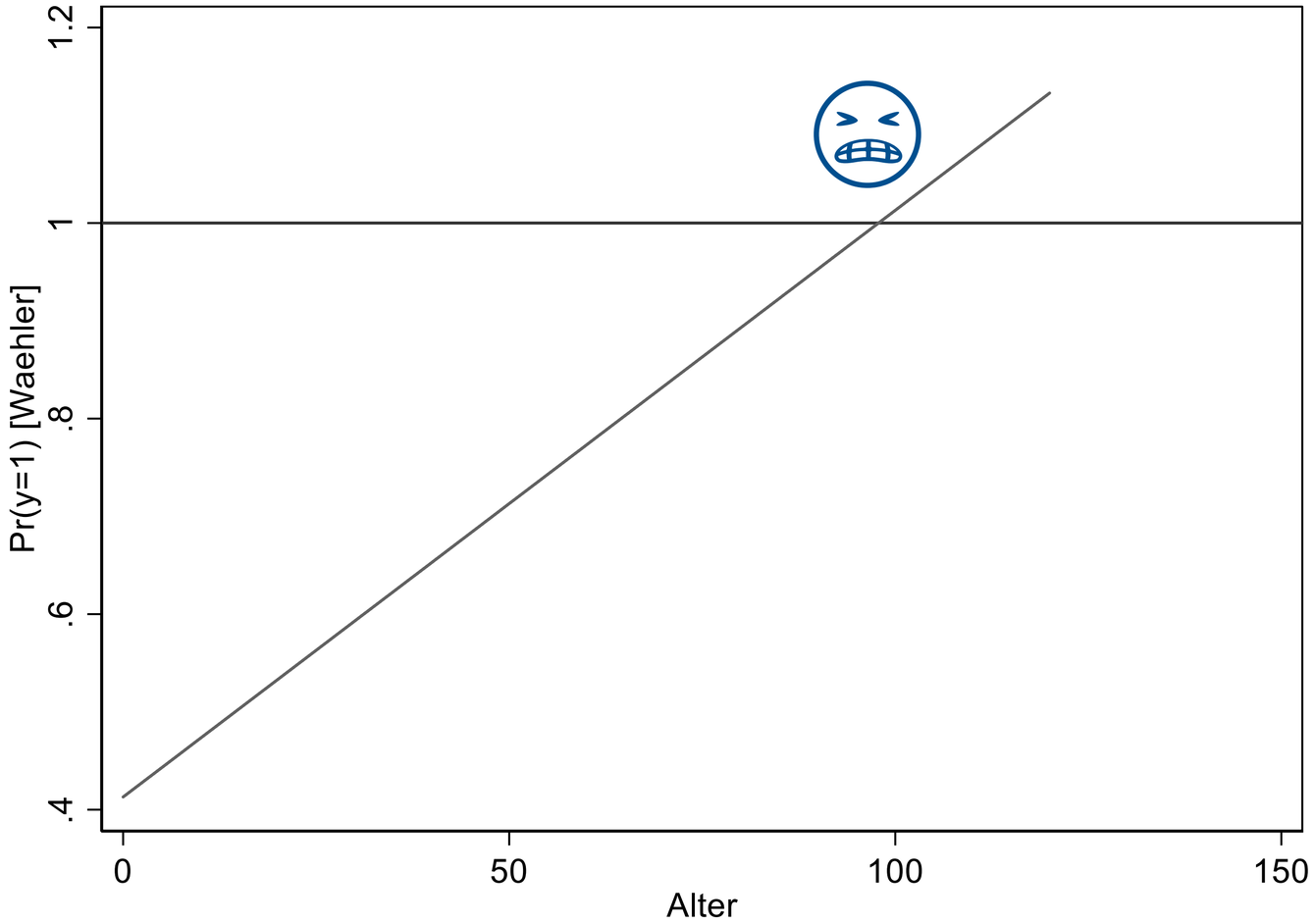
- So far we treated our outcomes as continuous
- Sometimes this assumption is at stake
- Dummy y?
- OLS → linear probability model (LPM)
  → `lm(y ~ x, data = data)`
- Maximum Likelihood (ML) → logistic regression
  → `glm(y ~ x, data = data, family="binomial")`
- Both part of base R

LINEAR PROBABILITY MODEL

OLS WITH DUMMY OUTCOMES

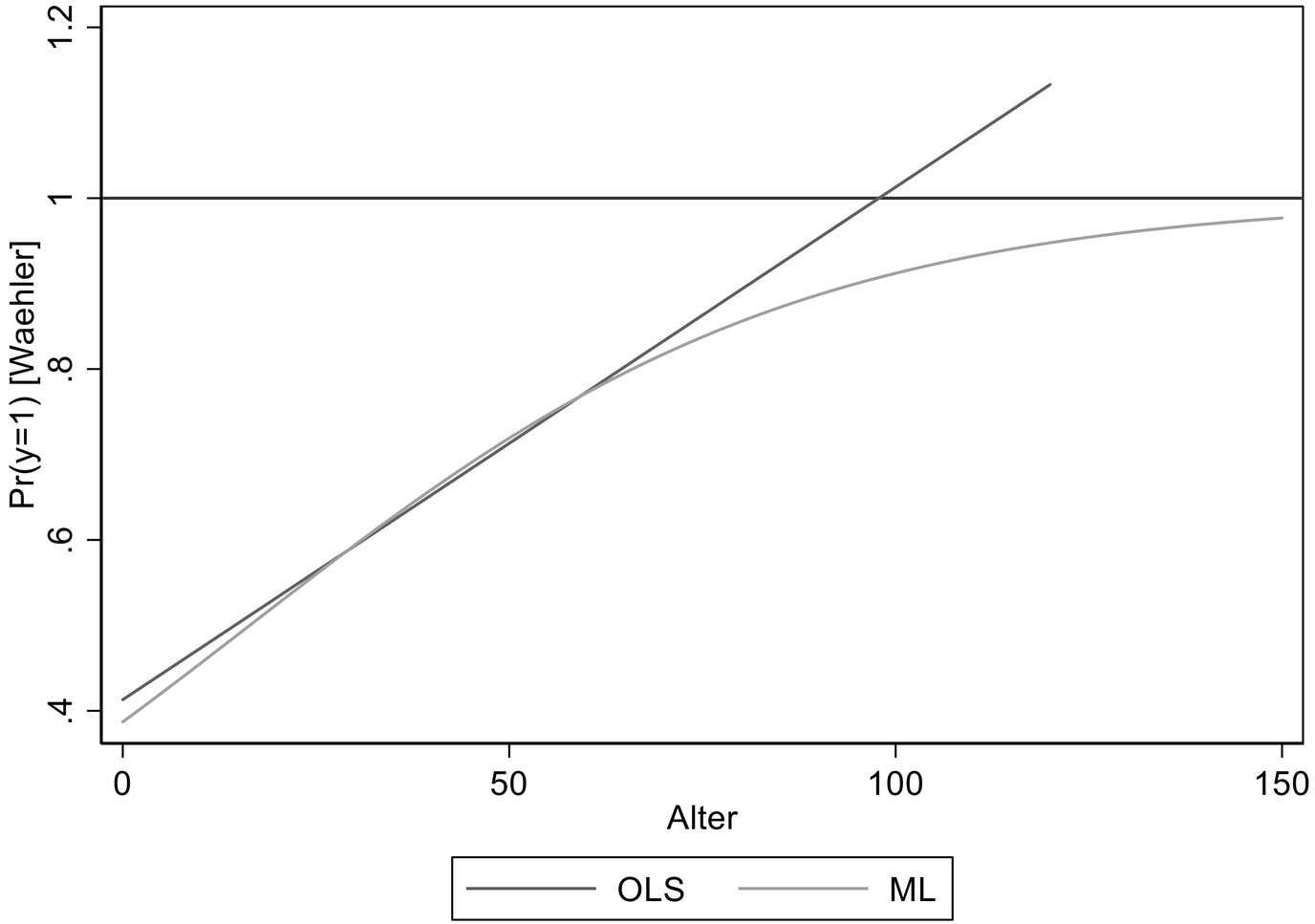
- The perhaps simplest way to deal with binary depended variables is to use OLS
- Benefits:
  - Effects linear and additive → relatively…
    - Straightforward interpretation
    - Fast computation
  - OLS BLUE if assumptions are met
- Interpretation: "*A one unit change in x is associated with a change in the probability to observe $y = 1$ of β_x percentage points*"
- *E. g.: "One more year of education increases the probability to vote by 4 percentage points"*
- But…

LINEAR PROBABILITY MODEL



LOGISTIC REGRESSION (PROLOGUE)

LPM VS LOGISTIC REGRESSION



REGRESSION RESULTS

- Logistic regression will always predict probabilities [0, 1]
- Estimates are logged odds (logits) or odds ratios
- … which are non-linear and multiplicative
- Odds are ratios of probabilities (Wahrscheinlichkeits-verhältnisse), odds ratios is the ratio of these ratios (Verhältnis von Wahrscheinlichkeitsverhältnissen)

ODDS & ODDS RATIOS

“WHAT ARE THE ODDS?”

	Group 1	Group 2	
Category 1	a	b	N(cat. 1)
Category 2	c	d	N(cat. 2)
	N(group 1)	N(group 2)	

- Odds: **Relations**, e. g. from cat. 1 to cat. 2:
  N(cat. 1) / N(cat. 2)
- Odds ratios: **Relations of relations**, e. g. from cat. 1 to cat. 2 for group 1 to group 2:
  (a/c) / (b/d)

INTERPRETATION OF ODDS RATIOS

	group 1	group 2
category 1	a	b
category 2	c	d

- Formula: $OR = \frac{a/c}{b/d} = \frac{a*d}{b*c}$

- Odds ratio > 1: odds of being in category 1 (compared to category 2) are higher in group 1 (compared to group 2)
- Odds ratio = 1: odds to be in category 1 are exactly equal between both groups
- Odds ratio < 1: odds are lower in group 1

INTERPRETATION OF ODDS RATIOS

- *"The odds / chances of being in category 1 rather than in category 2 are **[OR] times** larger (smaller) for group 1 than for group 2."*
- Alternatively: *"The odds /chances of being in category 1 rather than in category 2 are **[(OR-1)*100] percent** larger (smaller) for group 1 than for group 2."*

NOT "The probability…"!

ODDS VS PROBABILITY

- Simple example: flipping a coin
- Odds of heads (vs. tails): **1**(:1)
- Probability of heads: 50 percent (or **0.5**)
- Odds are the ratio of the probability to its counter probability: $odds(heads) = \frac{probability(heads)}{1-probability(heads)}$

EXAMPLE: VOTING

Socio-economic status (x)	Voted in last election? (y)		
	Yes (b1)	No (b2)	Total
High (a1)	**4,854(h11)**	**1,026 (h12)**	5,880 (h1.)
Low (a2)	**3,742 (h21)**	**2,683 (h22)**	6,425 (h2.)
Total	8,596 (h.1)	3,709 (h.2)	12,305 (n)

$$O(vote, not\ vote|high\ SES) = \frac{vote_{high\ SES}}{no\ vote_{high\ SES}}$$

$$O(vote, not\ vote|low\ SES) = \frac{vote_{low\ SES}}{vote_{low\ SES}}$$

$$OR = \frac{Odds(vote, not\ vote|high\ SES)}{Odds(vote, not\ vote|low\ SES)}$$

$$O(b1, b2|X = a_1) = \frac{h_{11}}{h_{12}} = \frac{4{,}854}{1{,}026} \approx \mathbf{4.73}$$

$$O(b1, b2|X = a_2) = \frac{h_{21}}{h_{22}} = \frac{3{,}742}{2{,}683} \approx \mathbf{1.39}$$

$$OR = \frac{O(b1, b2|X = a_1)}{O(b1, b2|X = a_2)} = \frac{h_{11}/h_{12}}{h_{21}/h_{22}} \approx \frac{4.73}{1.39} \approx \mathbf{3.4}$$

LOGISTIC REGRESSION

MAXIMUM LIKELIHOOD ESTIMATOR

- Wanted: Model of probability of observing $y = 1$
- Problem:
  - Observed data are values of y (1 & 0) and x
  - But not the probability
- Maximum Likelihood (ML) assumes that the observed y is a function of an underlying latent variable
- Strictly speaking, logistic regression models cannot be compared because the latent variable is re-scaled for each new model (see Breen et al. 2018)

OPTIMIZATION OF ML (SIMPLIFIED)

- ML tries different combinations of values for all parameters in the model
- Choses values which maximize probability of observing the given sample of y (and x) values

REGRESSION FOR PROBABILITY

- Modelling the probability to observe $y = 1$ (given x) → logistic distribution function:

$$\mathrm{pr}(y = 1|x) = \frac{e^{\beta_0+\beta_1 x_1}}{1 + e^{\beta_0+\beta_1 x_1}} \Leftrightarrow \ln\left(\frac{\mathrm{pr}(y = 1|x)}{1 - \mathrm{pr}(y = 1|x)}\right) = \beta_0 + \beta_1 x_1$$

- Non-linear model
- Coefficients are changes in the odds (odds ratios), respectively changes in the logged odds (logits)
- Note the missing error term

INTERPRETATION OF ML COEFFICIENTS

FOUR WAYS TO PRESENT RESULTS

I. Estimate the model using logistic distribution link function and report *logits*

$$\ln\left(\frac{\mathrm{pr}(y=1|x)}{1-\mathrm{pr}(y=1|x)}\right)=\beta_0+\beta_1 x_1$$

- Problem: almost impossible to understand intuitively

II. Transform coefficients from logits to *odds*

$$\frac{\mathrm{pr}(y=1|x)}{1-\mathrm{pr}(y=1|x)}=e^{\beta_0+\beta_1 x_1}=e^{\beta_0}*e^{\beta_1 x_1}$$

- Problems:
  - Effects non-linear & multiplicative → depend on values of all variables
  - Really more intuitive? (*not* probability but ratios of probability ratios!)

FOUR WAYS TO PRESENT RESULTS

III. Predict *average change* in probability of observing $y = 1$ as x increases by one unit, when all other variables are…

- Like observed in the data (*average marginal effect,* AME)
- At their means (*marginal effect at mean*)
- Yields similar coefficients of LPMs

IV. Predict probabilities for different combinations of certain values of all explanatory variables

$$\mathrm{pr}(y = 1|x) = \frac{e^{\beta_0+\beta_1 x_1}}{1 + e^{\beta_0+\beta_1 x_1}} = \frac{e^{\beta_0} * e^{\beta_1 x_1}}{1 + e^{\beta_0} * e^{\beta_1 x_1}}$$

- Problem: values must be chosen by researcher; with continuous predictors infinite possible combinations

- Substantively, all four ways contain the same information

LOGITS

- Default coefficients in, e. g., R: natural logarithm of odds of $y = 1$ (also called *logged odds* or *logits*)

$$\ln\left(\frac{\mathrm{pr}(y=1|x)}{1-\mathrm{pr}(y=1|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

  - $\beta = 0$: no differences
  - $\beta > 0$: positive relationship
  - $\beta < 0$: negative relationship

- Predicted *absolute* change in logits of $y = 1$ for a one unit increase of x

EXAMPLE: LOGITS

$$\ln\left(\frac{\mathrm{pr}(y=1|x)}{1-\mathrm{pr}(y=1|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

```
> summary(glm(voting ~ agea, data = data, family="binomial"(link = "logit")))

Call:
glm(formula = voting ~ agea, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6593  -1.3127   0.8550   0.9889   1.1710

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.208080   0.128682  -1.617    0.106
agea         0.014872   0.002547   5.839 5.26e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- β_1: "*Each additional year of age increases the logged odds of voting by 0.015*"
- Constant effect for all ages because logged odds are additive
- β_0: *log odds* of y for age 0
- Safe use: Focus on direction of effect (positive vs. negative) and statistical significance

ODDS RATIOS

- *"Change in the logged odds"*?!
- Exponential form transforms logits into ORs

$$\frac{\mathrm{pr}(y=1|x)}{1-\mathrm{pr}(y=1|x)} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k} = e^{\beta_0} * e^{\beta_1 x_1} * \ldots * e^{\beta_k x_k}$$

- *Non-linear* and *multiplicative* coefficients
  - *Relative* change by **factor** in the odds of $y = 1$ as x increases one unit
  - $\beta = 1$: no difference
  - $\beta > 1$: chances of observing $y = 1$ higher for $x + 1$ compared to x (positive relationship)
  - $\beta < 1$: chances of observing $y = 1$ lower for $x + 1$ compared to x (negative relationship)

EXAMPLE: ODDS RATIOS

$$\frac{\mathrm{pr}(y=1|x)}{1-\mathrm{pr}(y=1|x)} = e^{\beta_0} * e^{\beta_1 x_1} * \cdots * e^{\beta_k x_k}$$

```
> exp(0.014872)
[1] 1.014983
```

- β_1: "*Each year increases the odds of voting (vs. not voting) by the* ***factor*** *of 1.015*"
- Same value for all ages because increase not in absolute units but in the relative factor
- β_0: Odds of voting for age 0

EXAMPLE: AVERAGE MARGINAL EFFECTS

```
> margins::margins(glm(voting ~ agea, data = data, family="binomial"(link = "logit")))
Average marginal effects
glm(formula = voting ~ agea, family = binomial(link = "logit"),     data = data)

     agea
 0.003433
```

- Average change in probability of $y = 1$ when x increases by one unit
- "*The probability that a person voted increases on average by about 0.3 percentage points with each year of age*"
- See: https://cran.r-project.org/web/packages/margins/vignettes/Introduction.html#Average_Marginal_Effects_and_Average_Partial_Effects

EXAMPLE: PREDICTED PROBABILITY

```
Call:
glm(formula = voting ~ agea, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6593  -1.3127   0.8550   0.9889   1.1710

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.208080   0.128682  -1.617    0.106
agea         0.014872   0.002547   5.839 5.26e-09 ***
```

- $pr(y = 1|x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$

→ $\Pr(waehler = 1|alter = 18) = \frac{e^{-0.208+0.0149*18}}{1+e^{-0.208+0.0149*18}} = 0.515$

→ $\Pr(waehler = 1|alter = 19) = \frac{e^{-0.208+0.0149*19}}{1+e^{-0.208+0.0149*19}} = 0.5188$

→ $\Pr(waehler = 1|alter = 40) = \frac{e^{-0.208+0.0149*40}}{1+e^{-0.208+0.0149*40}} = 0.5958$

→ $\Pr(waehler = 1|alter = 41) = \frac{e^{-0.208+0.0149*41}}{1+e^{-0.208+0.0149*41}} = 0.5994$

EXAMPLE: PREDICTED PROBABILITIES

- $Pr(waehler = 1|alter = 18) = 0.515$
- $Pr(waehler = 1|alter = 19) = 0.5188$

$\Delta = -0.0038$

- $Pr(waehler = 1|alter = 40) = 0.5958$
- $Pr(waehler = 1|alter = 41) = 0.5994$

$\Delta = -0.0036$

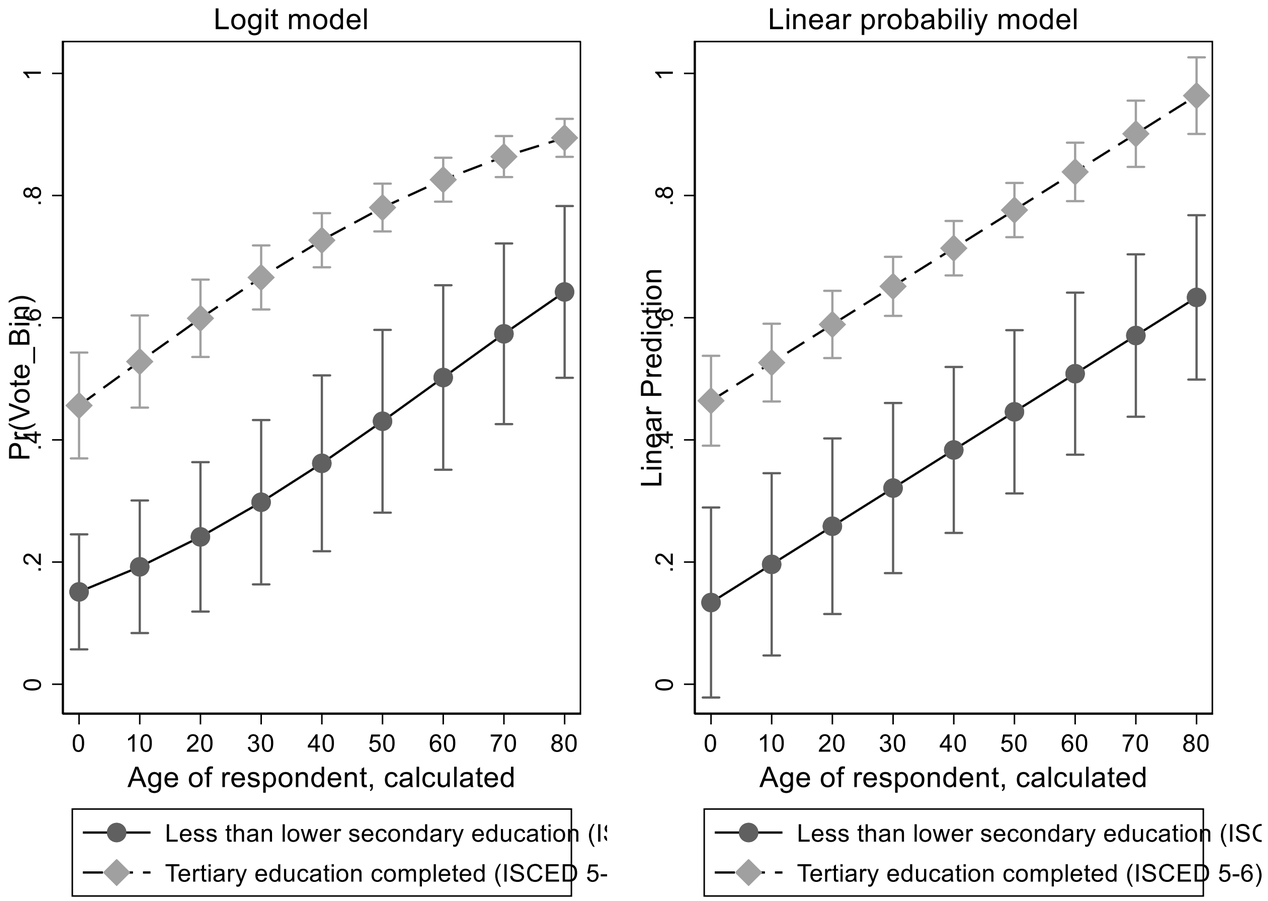
- Each predicted probability is conditional on values
  - Of x itself
  - On the values of all other explanatory variables
  - Again: underlying model is non-linear & multiplicative

LINEAR PROBABILITY MODEL VS LOGISTIC REGRESSION

THINKING THE UNTHINKABLE?

- Coefficients of LPM easily interpretable, no easy interpretation for logit coefficients
- … making it hard to interpret actual effect sizes, not just statistical significance (Gomila 2021)
- "*Using LPM is almost unthinkable in sociology, while it is common in economics*" (Mood 2010: 78)
- LPM provides an easy and straightforward way to estimate average effects
- AME do the same for the logistic case, but "*deriving AME from logistic regression is just a complicated detour*" (Mood 2010: 78)

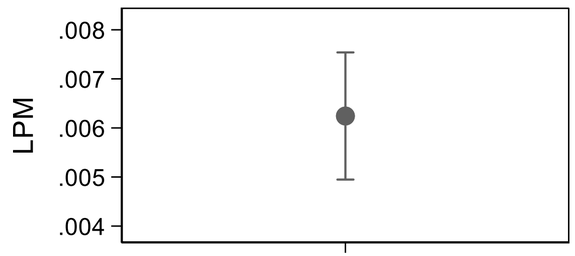
REGRESSION SLOPES FOR AGE EFFECT



SIZE OF AGE EFFECT FOR DIFFERENT VALUES OF CONTROL VARIABLES

Effect of age on probability to vote
Covariates at ...

Minima Maxima



Based on ESS4 (Latvia only)
Model controlling for edulvla, stfdem, euftf, imueclt

CHOOSING BETWEEN LPM AND LOGISTIC MODEL

- Chose the estimator based on theoretical and empirical considerations
- Run LPM and logistic models and compare results
  - Similar conclusions: why not use LPM? (KISS principle)
  - Different conclusions: perhaps use logistic regression
- But: Predicted probabilities should be [0, 1]
- Does the research question include more "extreme" values?
  - E. g. focus on the "oldest old"
  - Effect sizes between logit and LPM may differ more for very high (or low) values
- Many zeros in outcome?
  - E. g. Moving between cities

LITERATURE

- Mood (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. European Sociological Review 26 (1): 67-82.
- Breen Karlson & Holm (2018). Interpreting and understanding logits, probits, and other nonlinear probability models. Annual Review of Sociology 44: 39-54.