Predictions and effects

Jonathan Dushoff, McMaster University

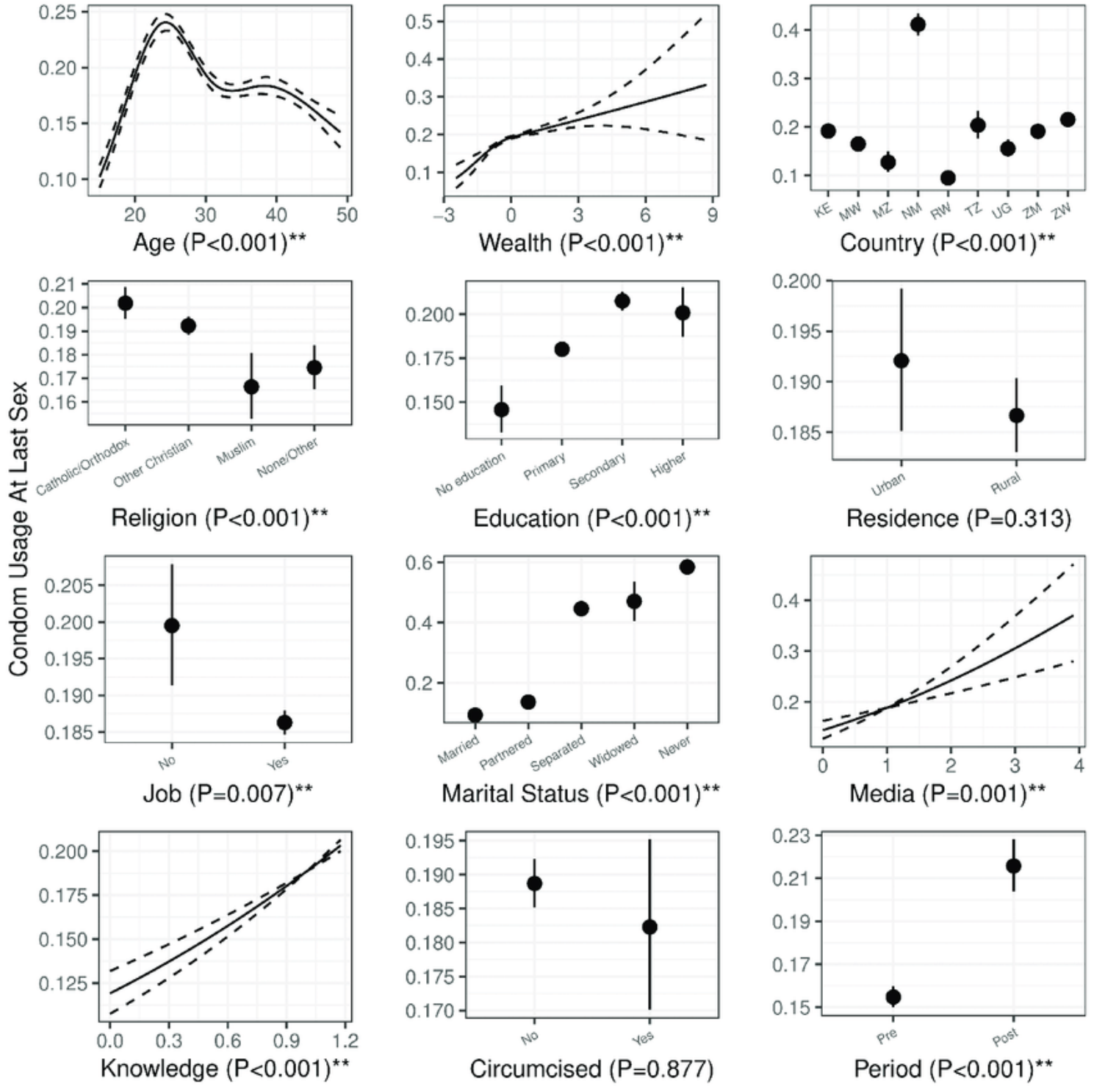
McMaster Statistics Seminar, March 2022

Thanks

- Bicko Cygu
- Ben Bolker
- Mike Li
- Chyun Shi

Visualizing results of complicated models

- What are the factors that predict?:
  - HIV risk
  - Tendency to circumcise sons and daughters
  - Access to clean water
- A variety of "socio-demographic" variables and complicated models


Condom Usage At Last Sex
Age (P<0.001)**
Wealth (P<0.001)**
Country (P<0.001)**
KE MW MZ NM RW TZ UG ZM ZW
Religion (P<0.001)**
Catholic/Orthodox
Other Christian
Muslim
None/Other
Education (P<0.001)**
No education
Primary
Secondary
Higher
Residence (P=0.313)
Urban
Rural
Job (P=0.007)**
No
Yes
Marital Status (P<0.001)**
Married
Partnered
Separated
Widowed
Never
Media (P=0.001)**
Knowledge (P<0.001)**
Circumcised (P=0.877)
No
Yes
Period (P<0.001)**
Pre
Post

Types of variables

- An "input variable" can be associated with one or more parameters
  - And therefore one or more "model variables"
- Gender typically has a single parameter (two categories)
- Religion or ethnicity typically has multiple parameters (more than two categories)
- Wealth or education *may* have a single parameter (linear response)
- Age typically has multiple parameters (polynomial or spline response)

Tables

- Tables are generally bad for communicating model results
- Hard to see patterns
- Focus attention inappropriately on statistical “significance”

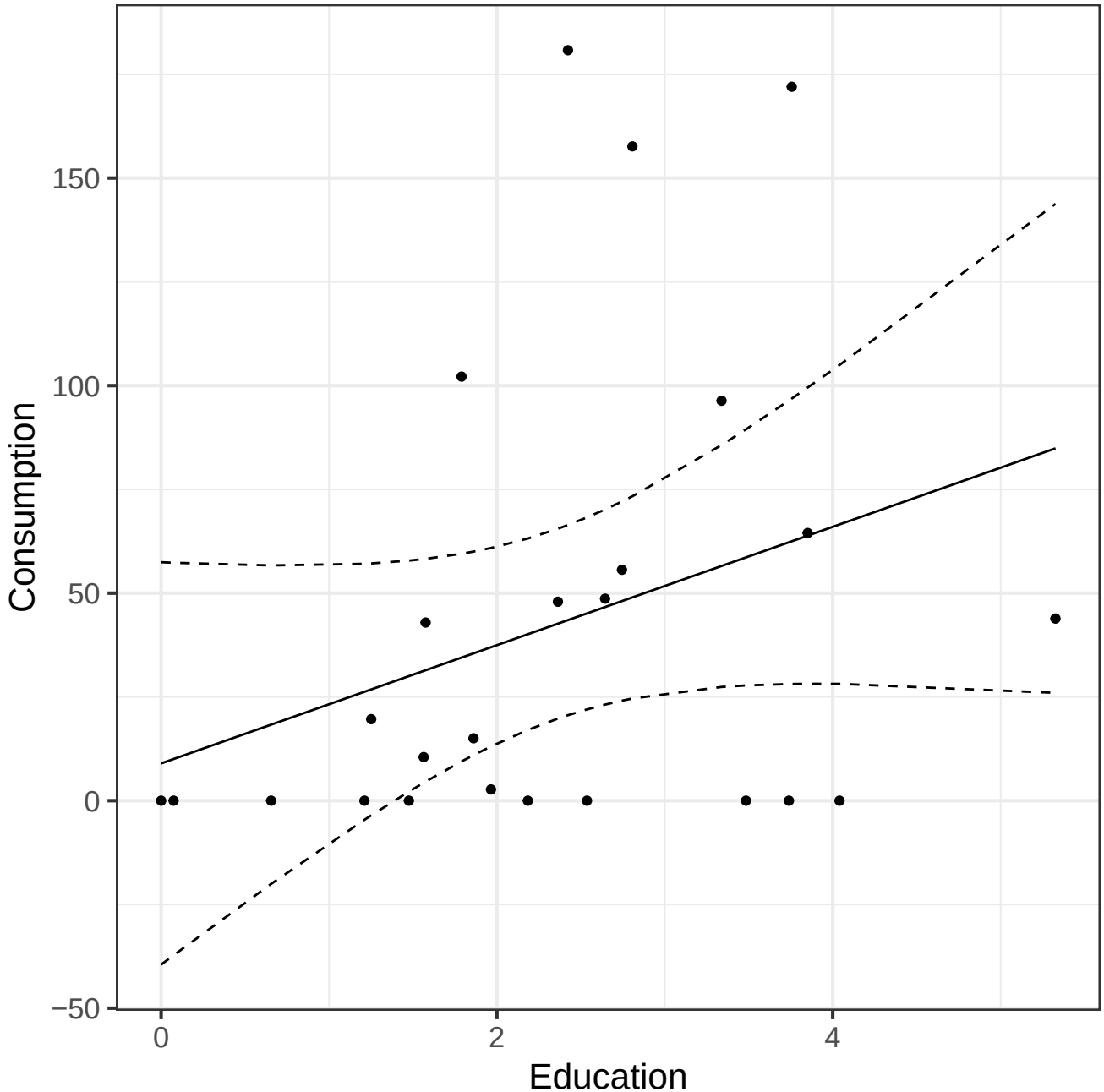
Coefficient plots

- Coefficient plots are a compact, effective way to communicate about single-parameter variables
  - Effects on the same axis should have the same units!
  - Normalize predictor variables under most circumstances
- They are not a good way to communicate about multi-parameter variables
  - Especially splines and polynomials

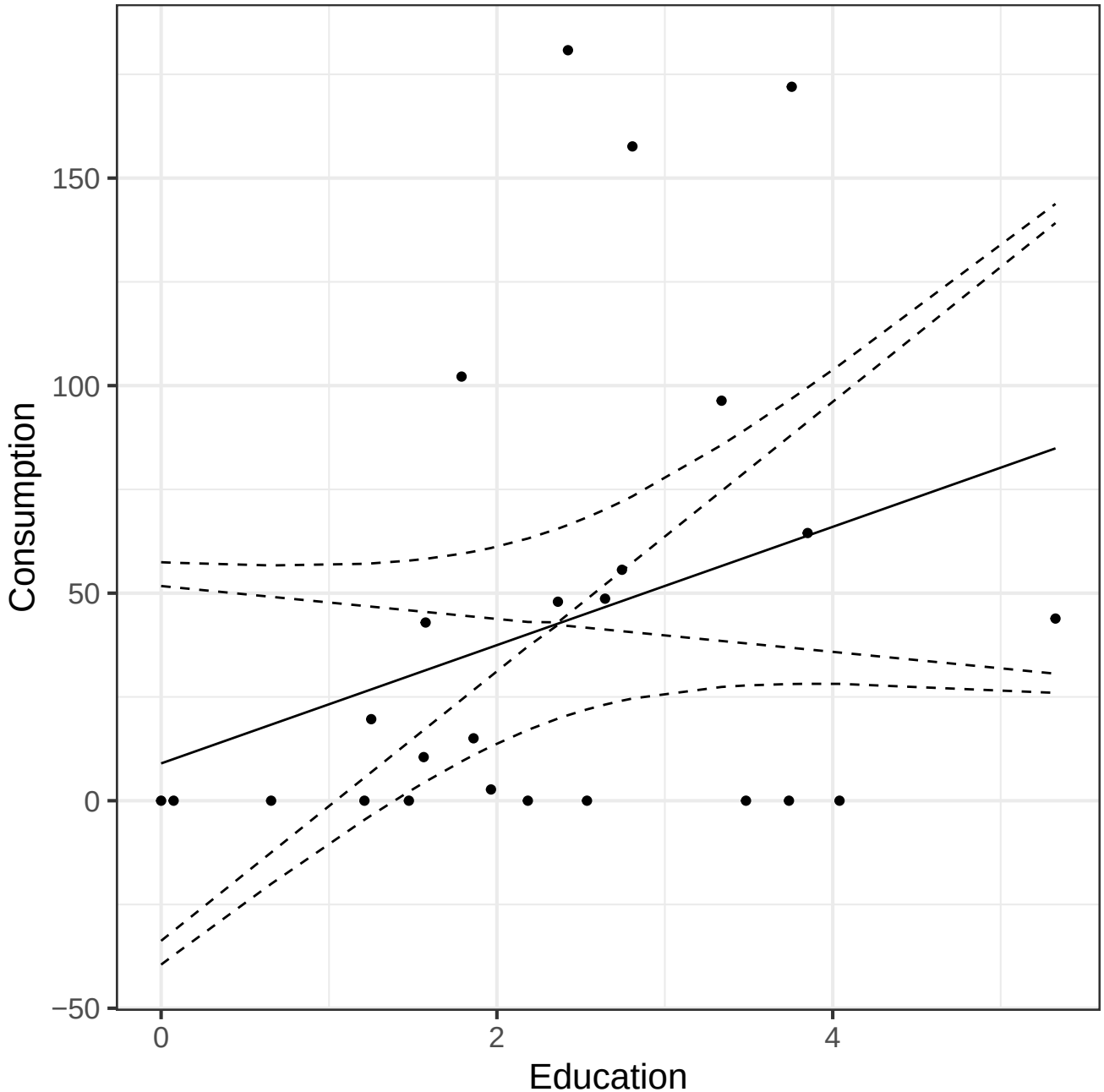
Effects plots

- An appealing way to visualize the results of model inference
- Particularly for multi-parameter variables
- Or possibly to aid understanding of generalized models
- Raise a number of interesting questions

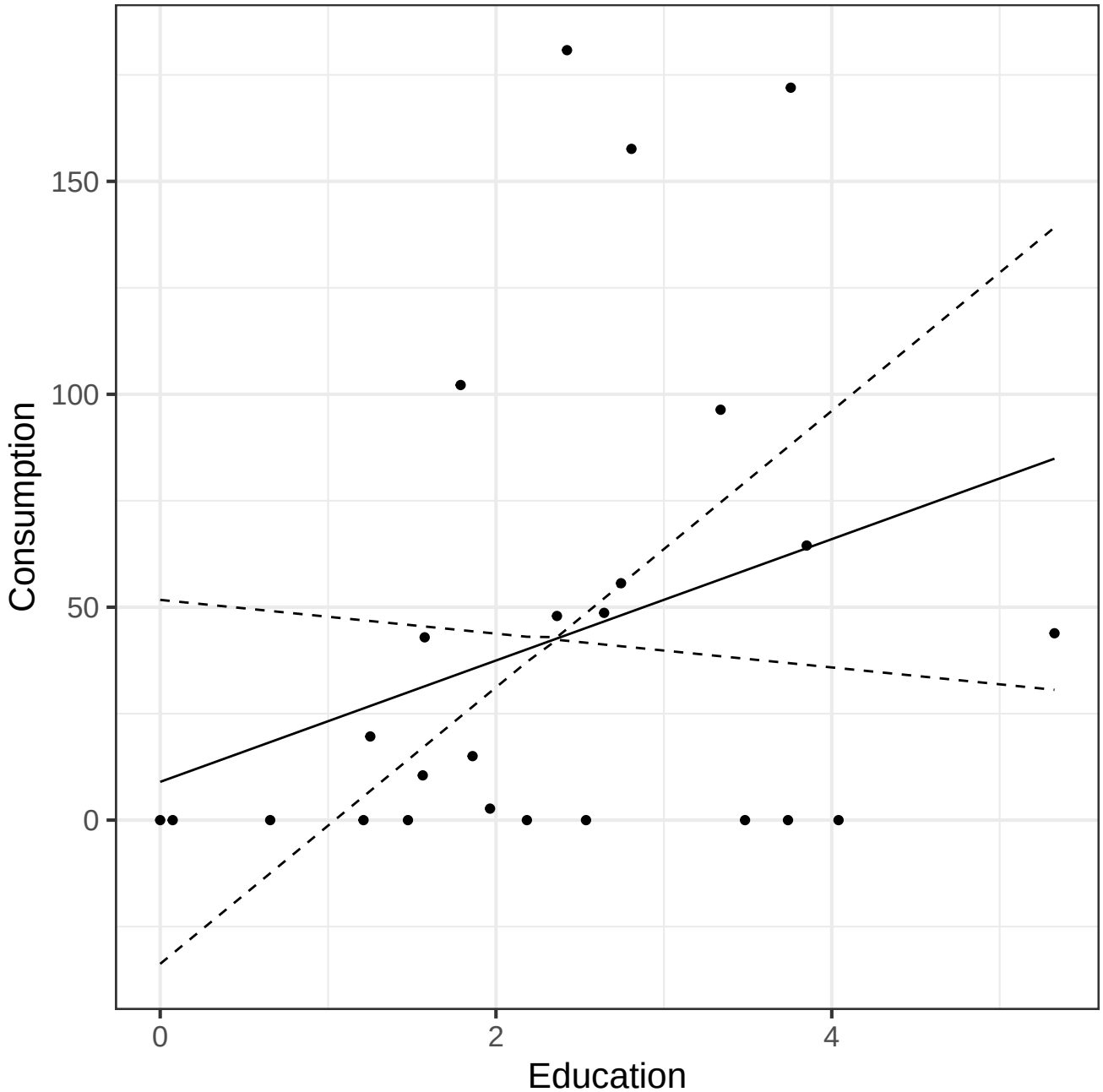
Predictions vs. effects



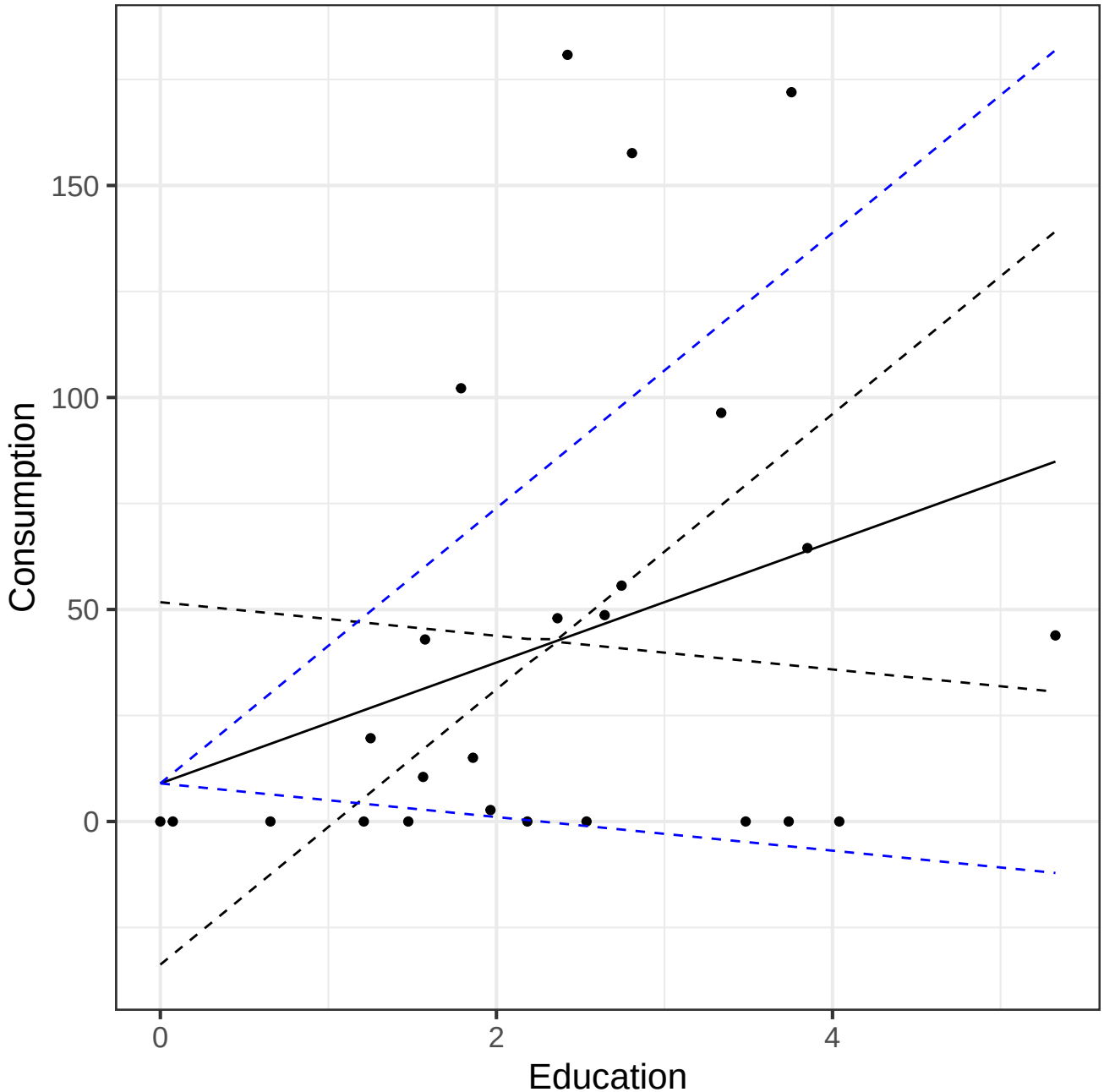
Predictions vs. effects



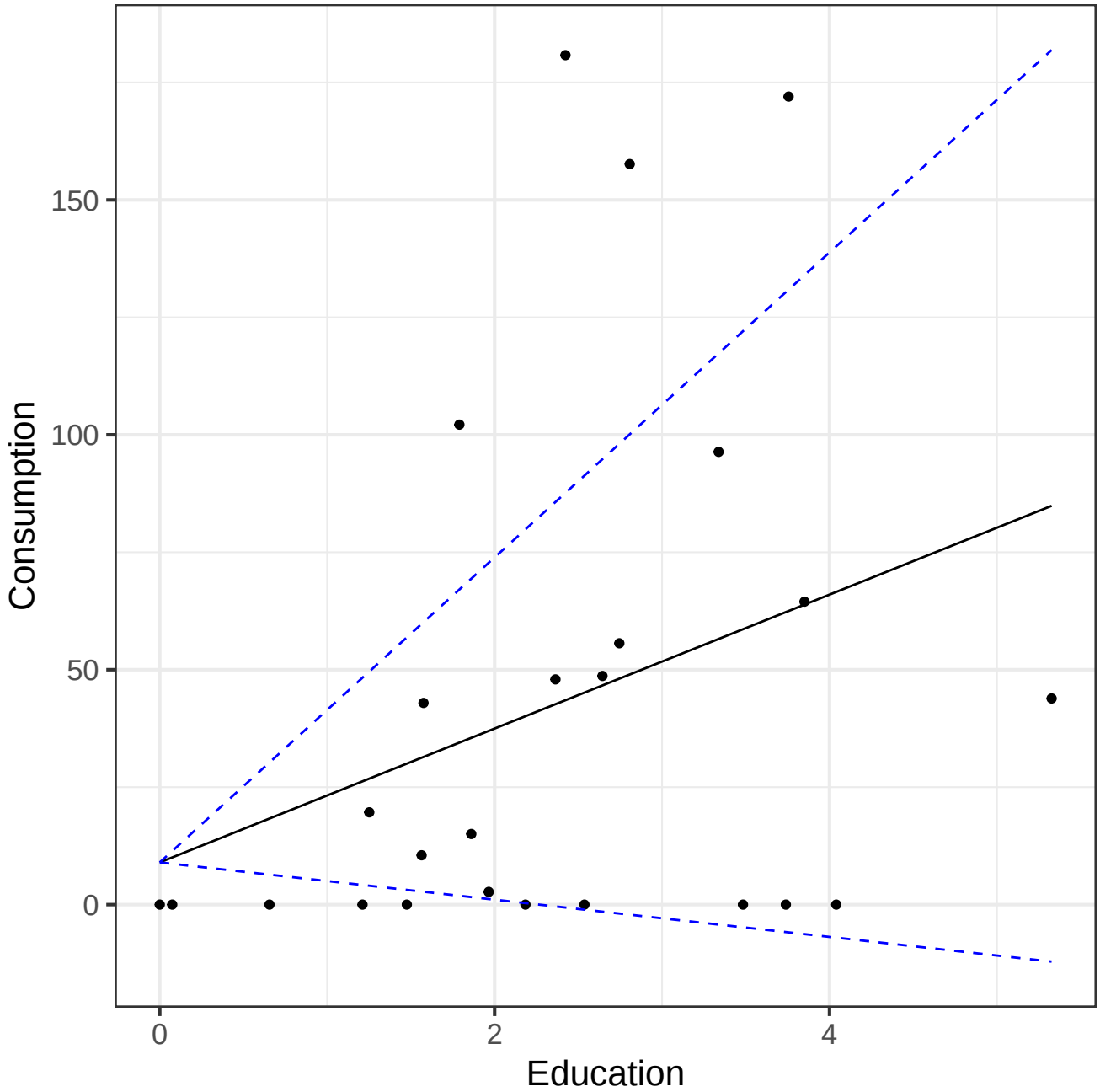
Anchor



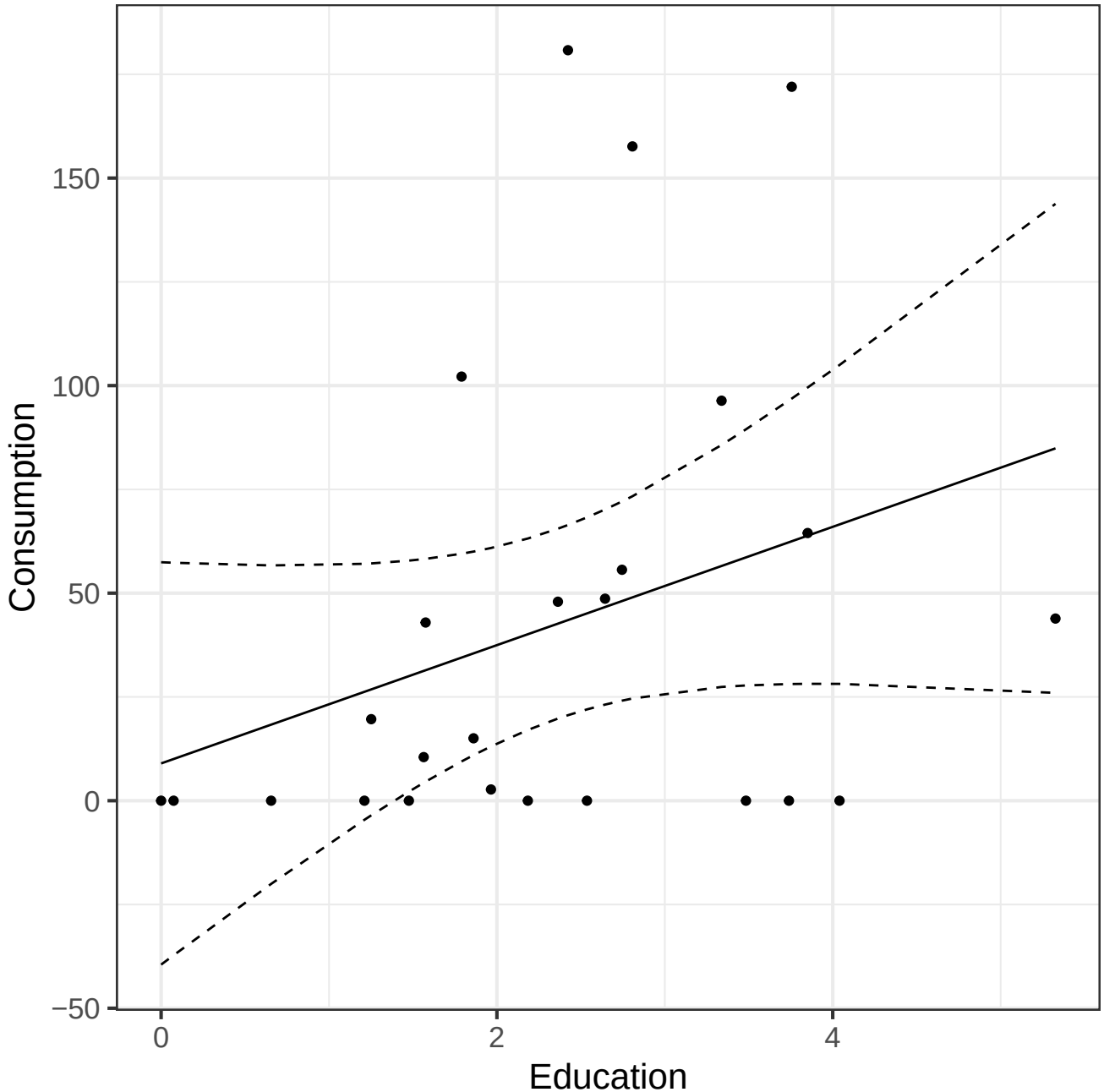
Anchor



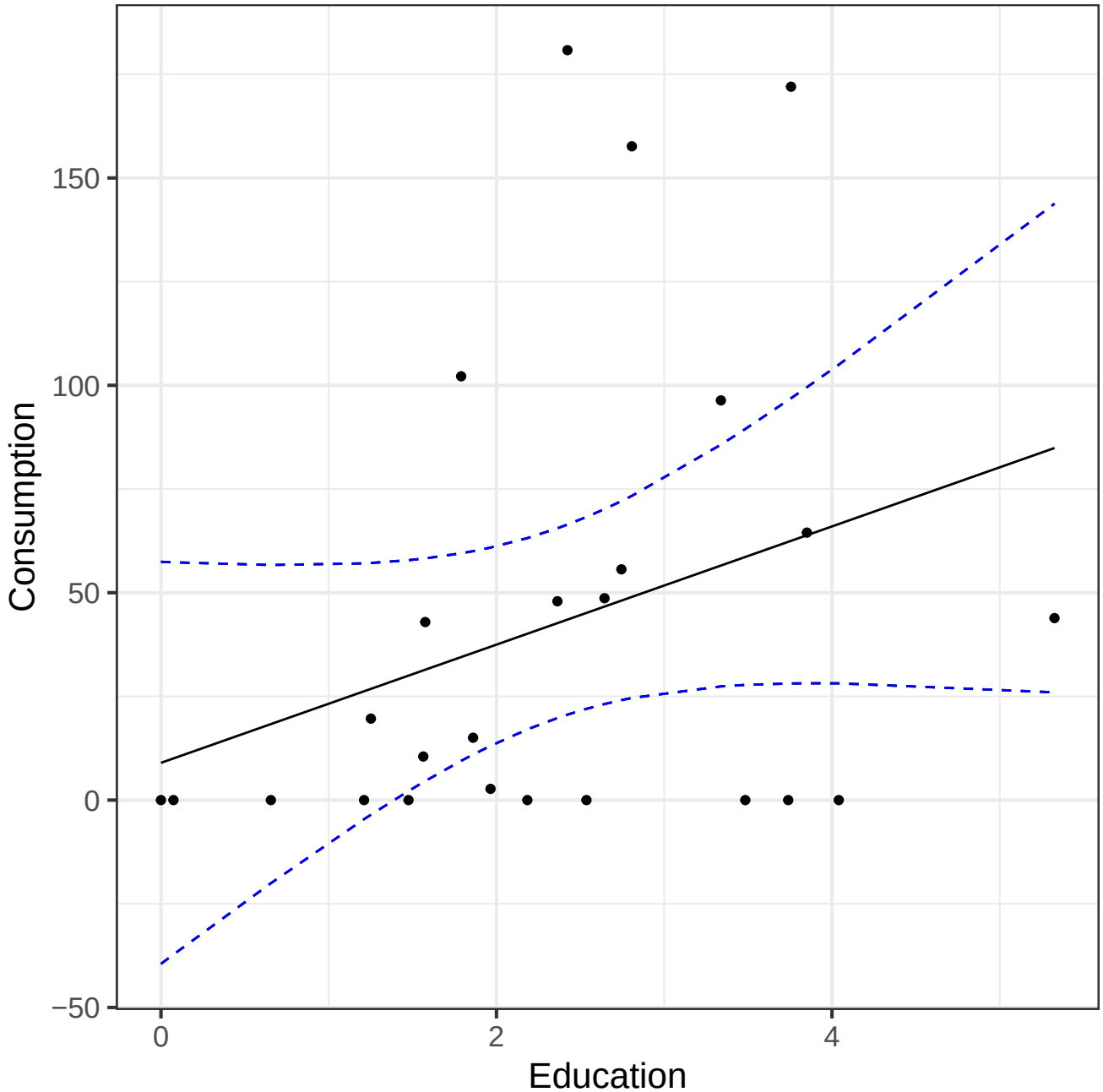
Anchor



No anchor

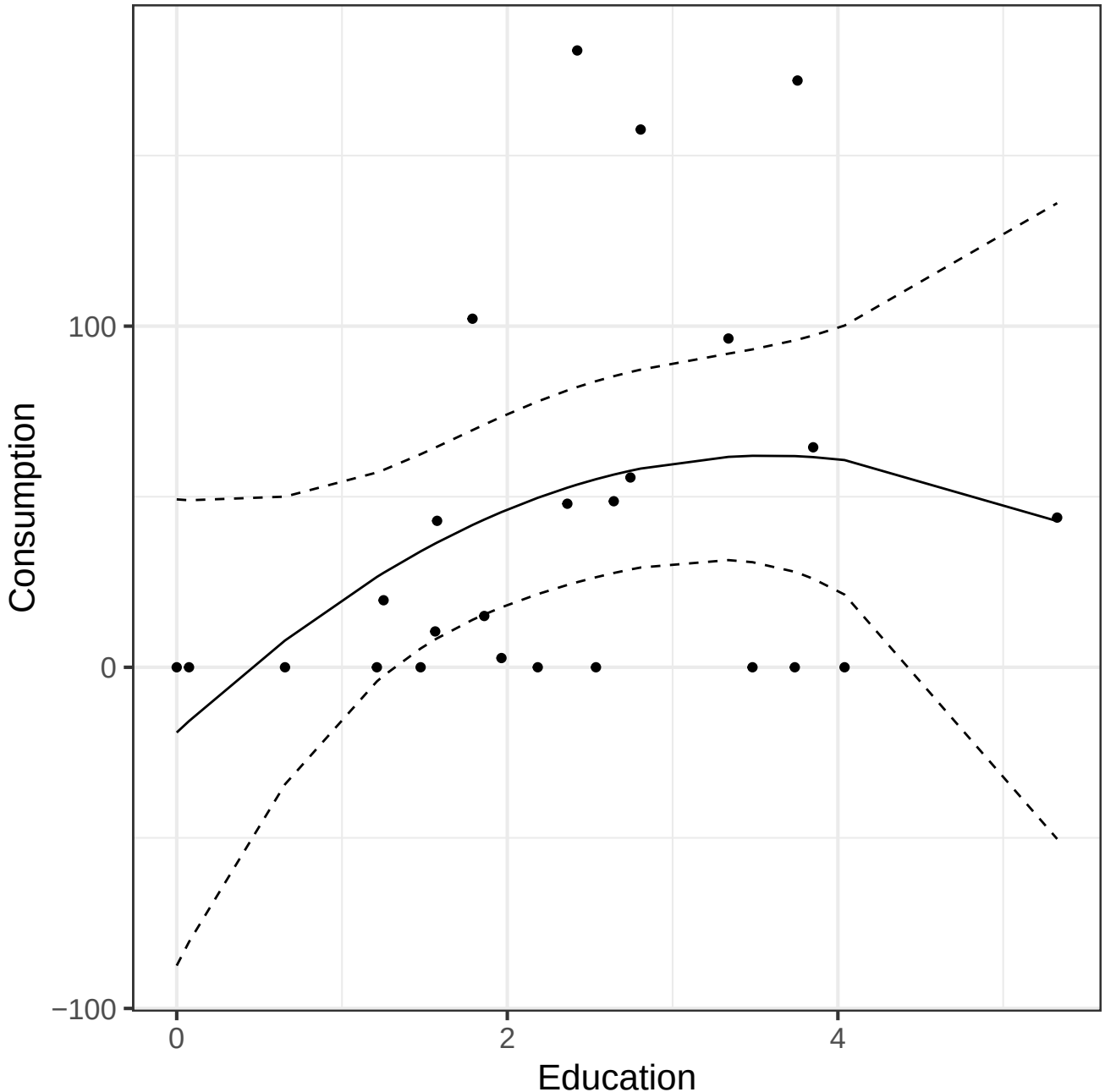


No anchor



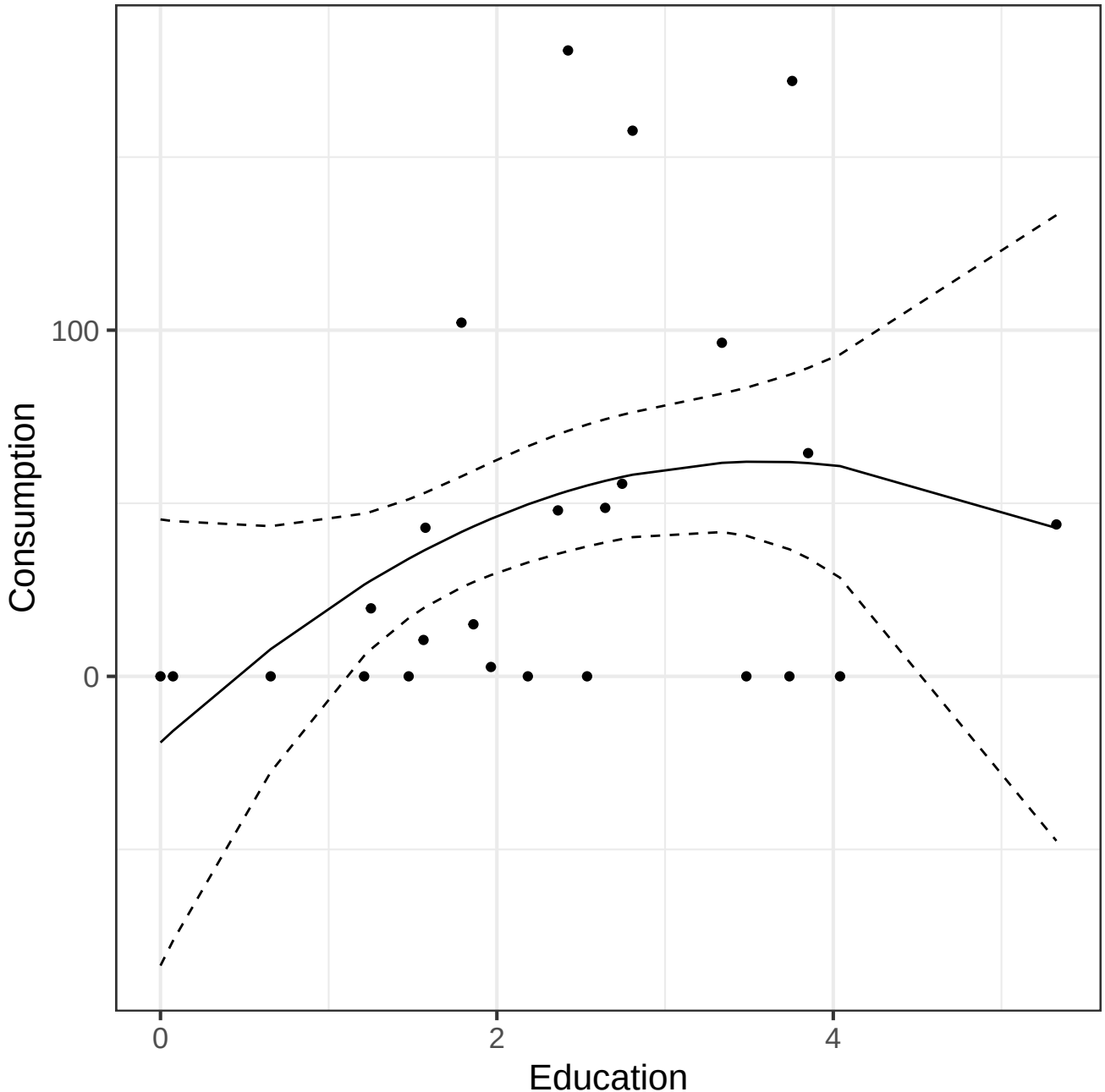
Multi-parameter predictor

Prediction

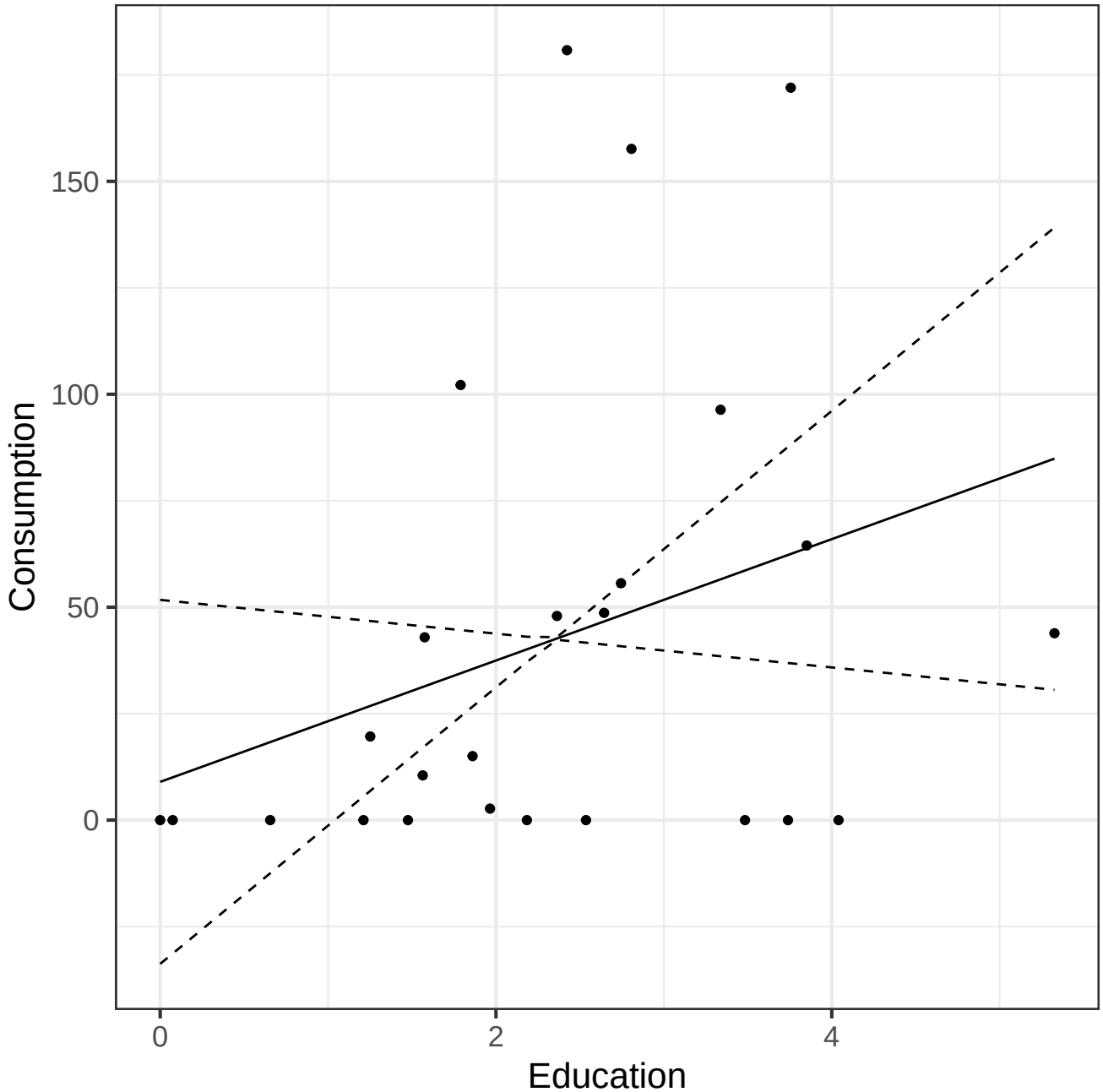


Multi-parameter predictor

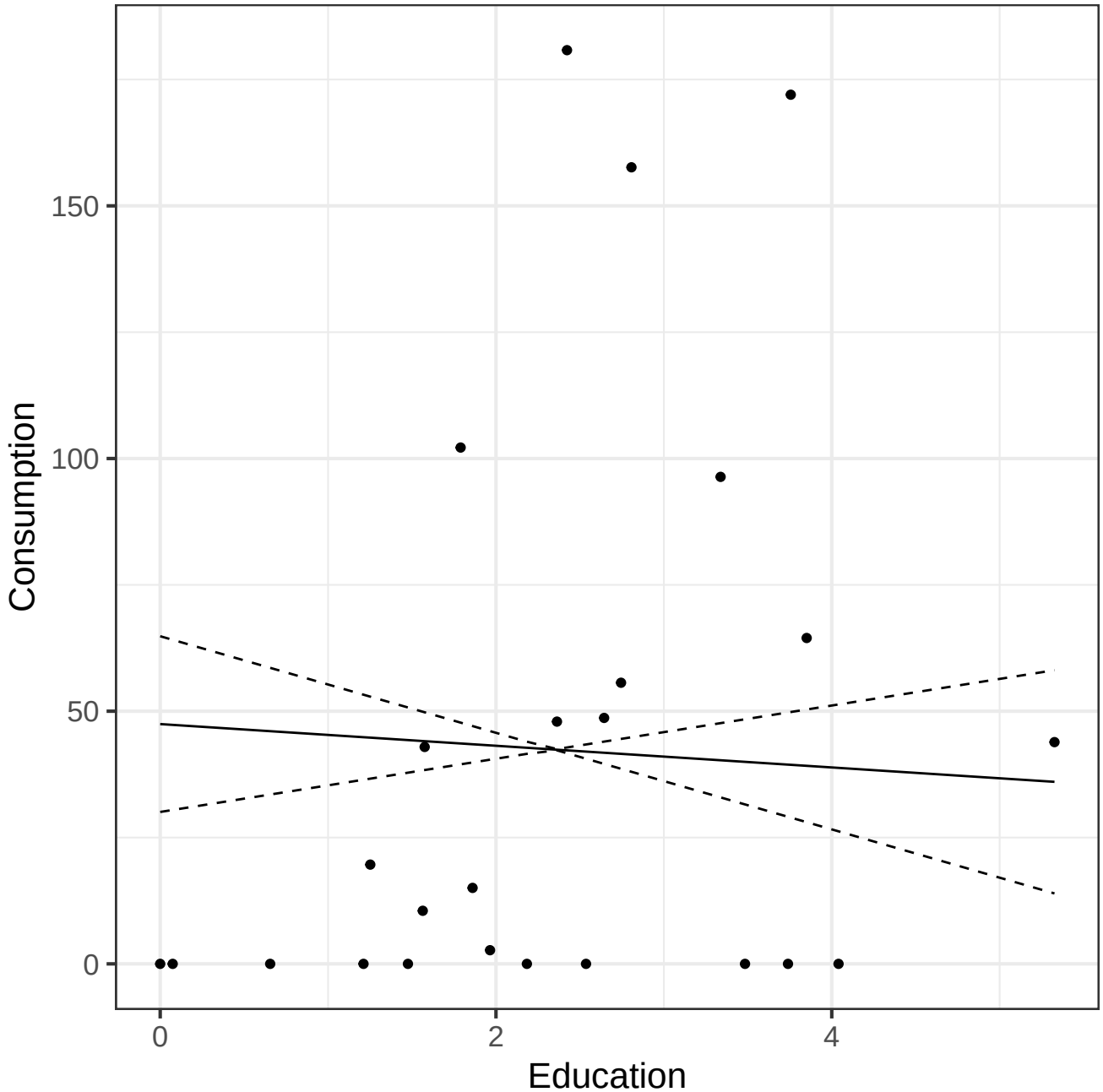
Effect



Univariate model



Multivariate model



The Model Center

- We define "model center" as the point at the mean of the *model* variables
  - Columns of the model matrix
- Centered columns are orthogonal to the intercept

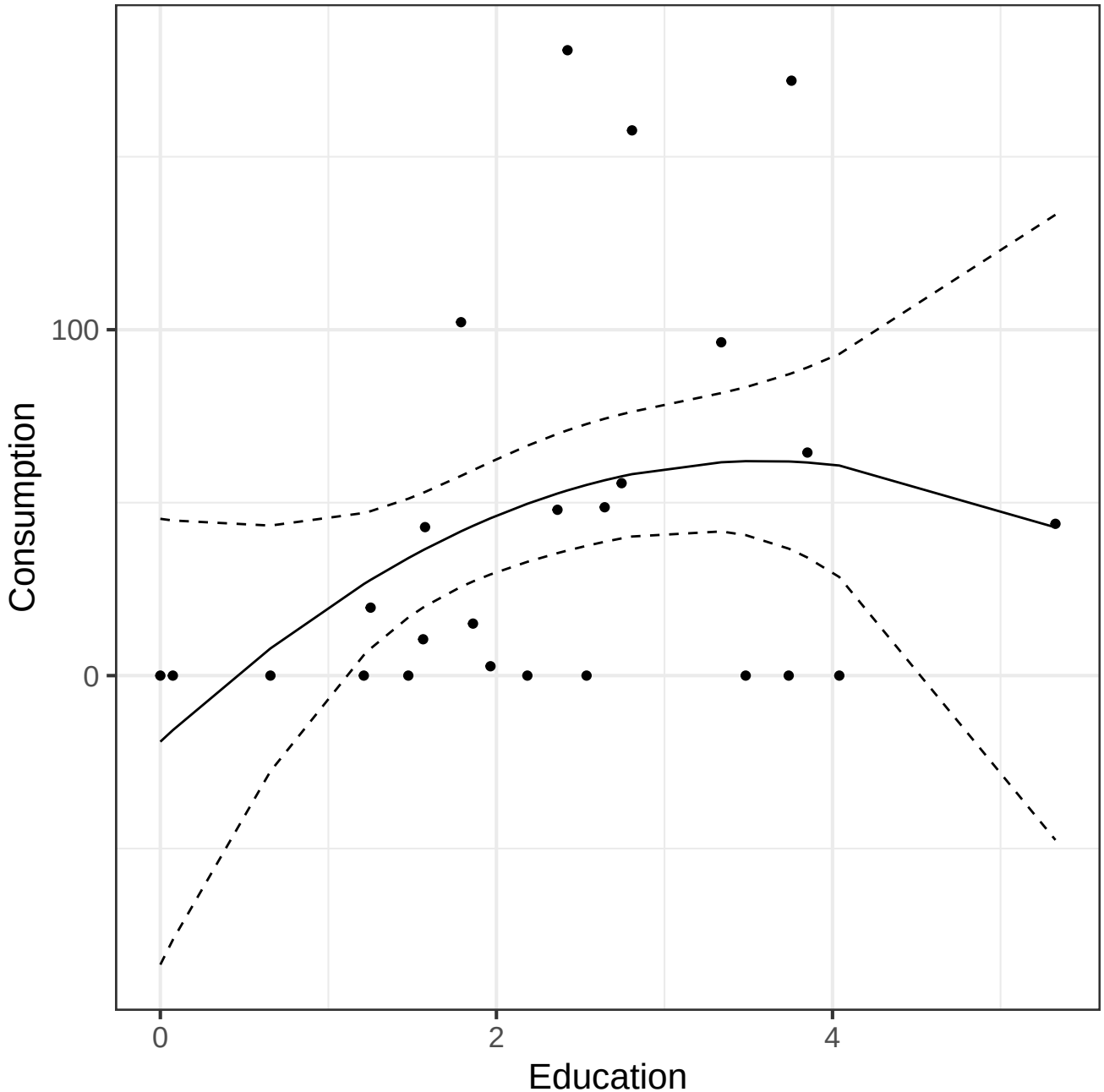
Proxies

- Some common tricks to stabilize models may be imperfect attempts to move toward the model center:
- Averaging input variables
  - Works perfectly for single-parameter variables
- Sum-to-zero contrasts for categories
  - Works perfectly for balanced designs
  - Or if we use weights

Anchor

- To calculate confidence intervals for (narrow-sense) effect plots, we need an *anchor*
  - Values for the *focal* statistical parameters defined as zero effect
- We argue that the model center should be the default
  - . . . but it's not the only sensible value

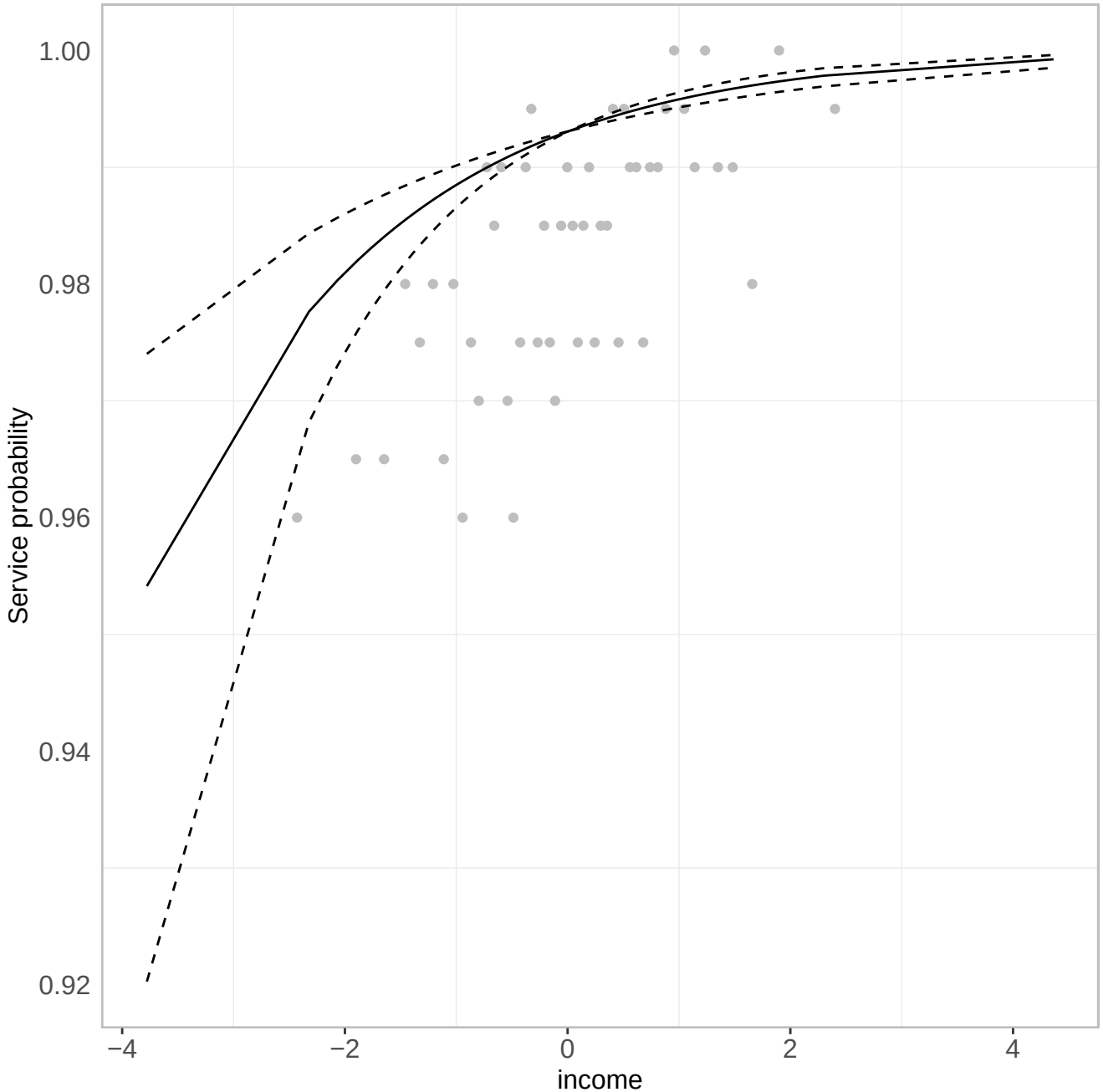
Center anchoring



Reference point

- In a multivariate model, to calculate any estimates for (any) effect plots, we need a reference point
  - Values to use for the *non-focal* statistical parameters
- We again argue that the model center should be the default
  - Random effects can basically be neglected (we will center them by accident)
  - ... but it's going to turn out to be complicated for generalized models

Generalized models



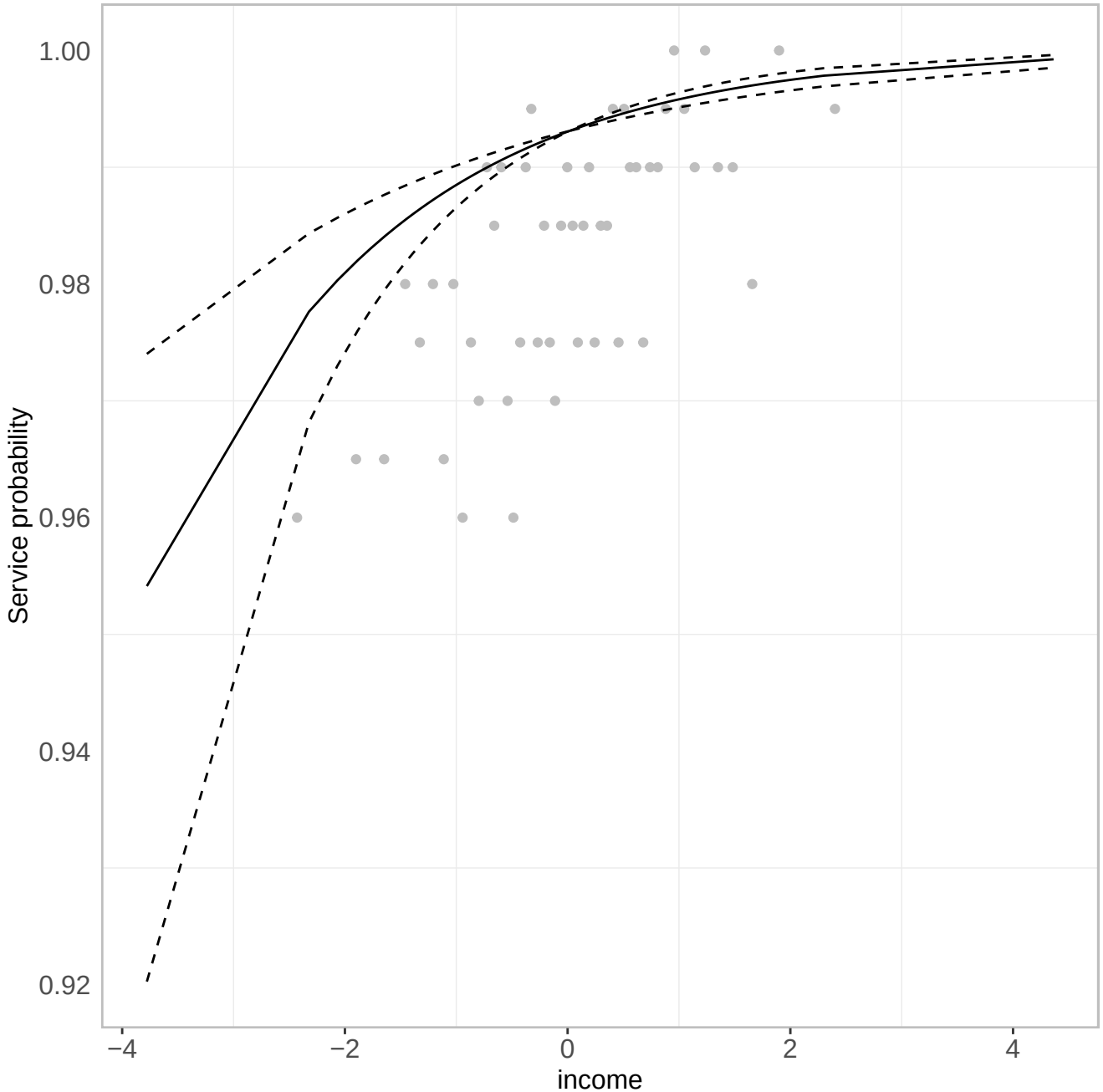
Non-linear averaging

- The model center is not as beautiful here!
  - We need to generalize our idea of reference point
  - i.e., the values we consider for non-focal predictors
- The prediction of the averaged population $\neq$ the average of the predictions
- More work needs to be done

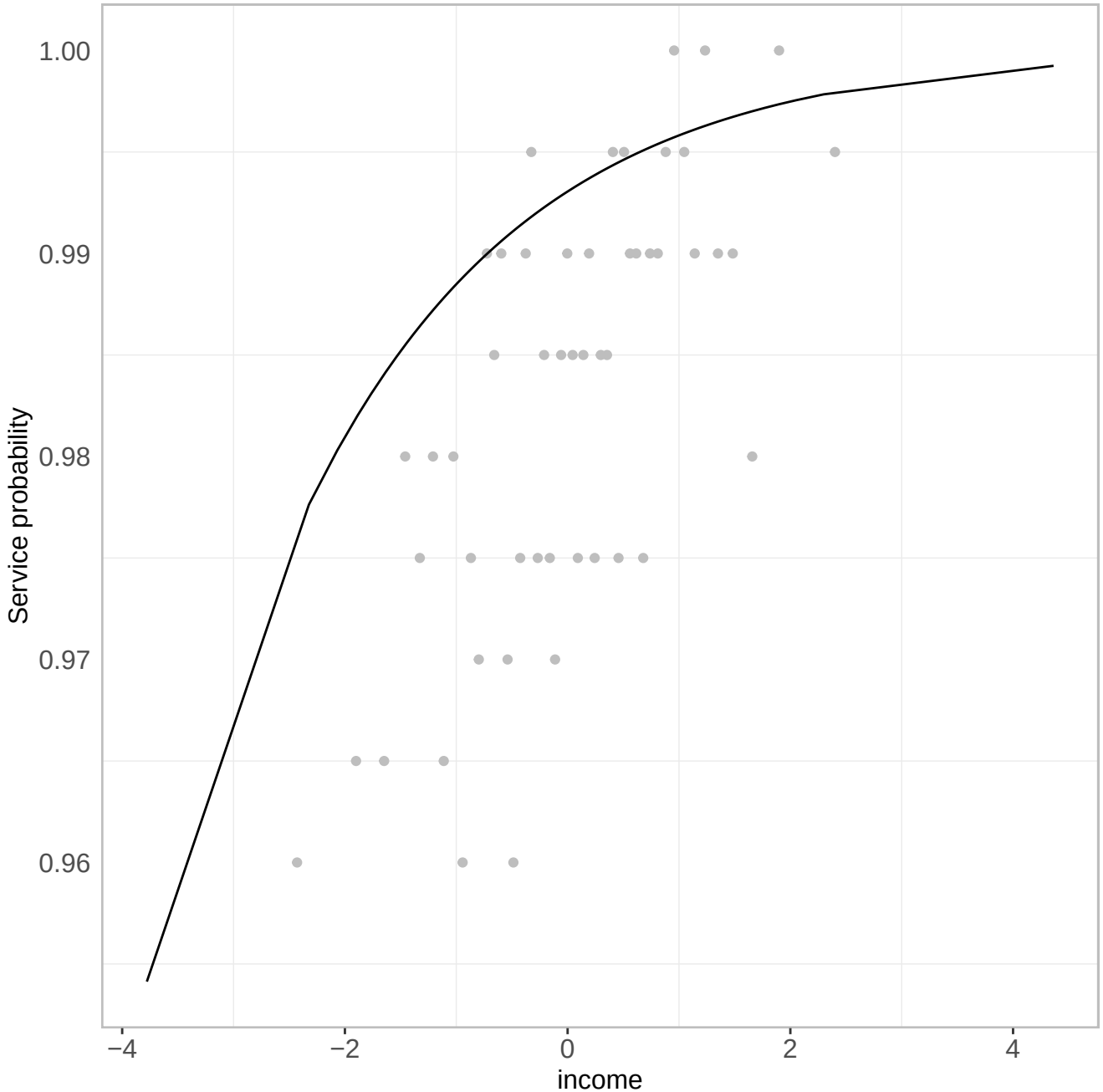
Bias correction

- Which biases should be corrected?
  - Use the linear model as a guide
- How to correct them?
  - Moment methods (Jensen)
  - Quantile-based methods
  - Population-based methods
    - Seem best; in some cases may be computationally demanding
    - Maybe solved by sampling

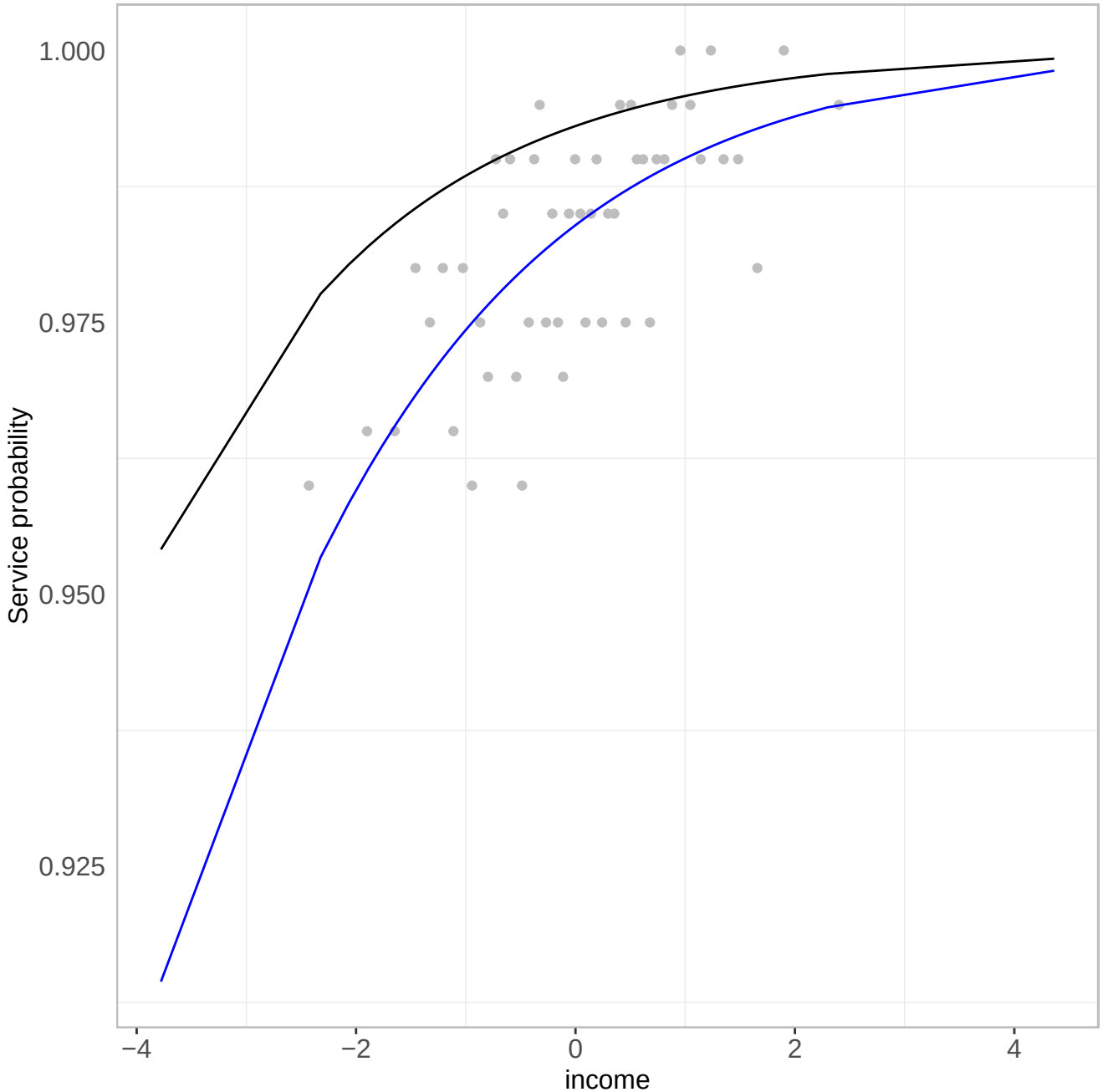
Naive use of model center



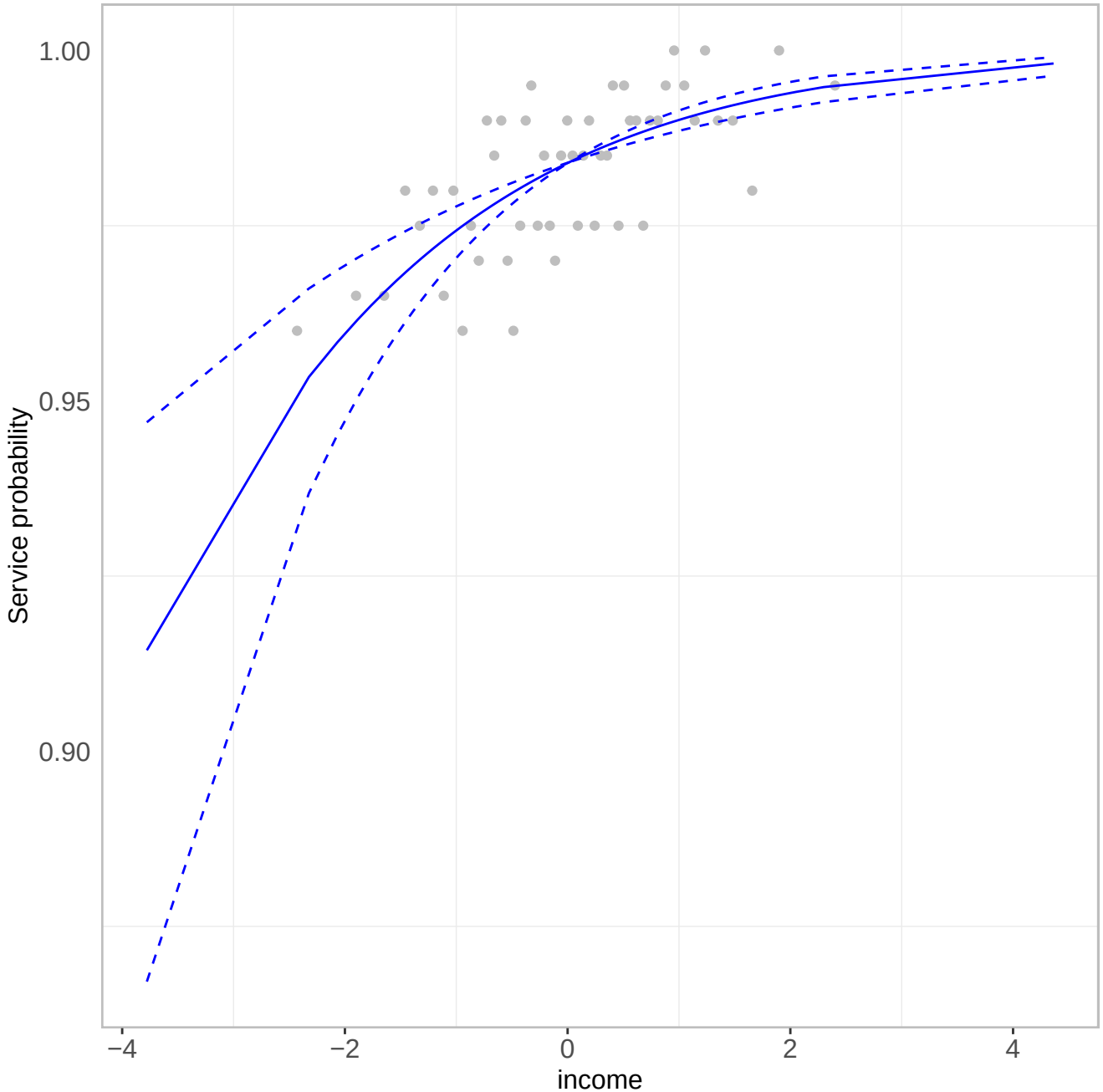
Population-based correction



Population-based correction



Population-based correction



Mixed models

- Averaging over random effects turns out to be pretty similar to averaging over fixed effects
  - In lm, you don't even need to think about it
  - In glm, you can correct using three methods discussed above

Work in progress

- Warning: alpha-level software!
- `https://github.com/mac-theobio/effects`

Thanks again

- Organizers
- Audience