

Survey data analysis  
Week 8:  
“Nonresponse and nonresponse weights”

© Peter Lugtig

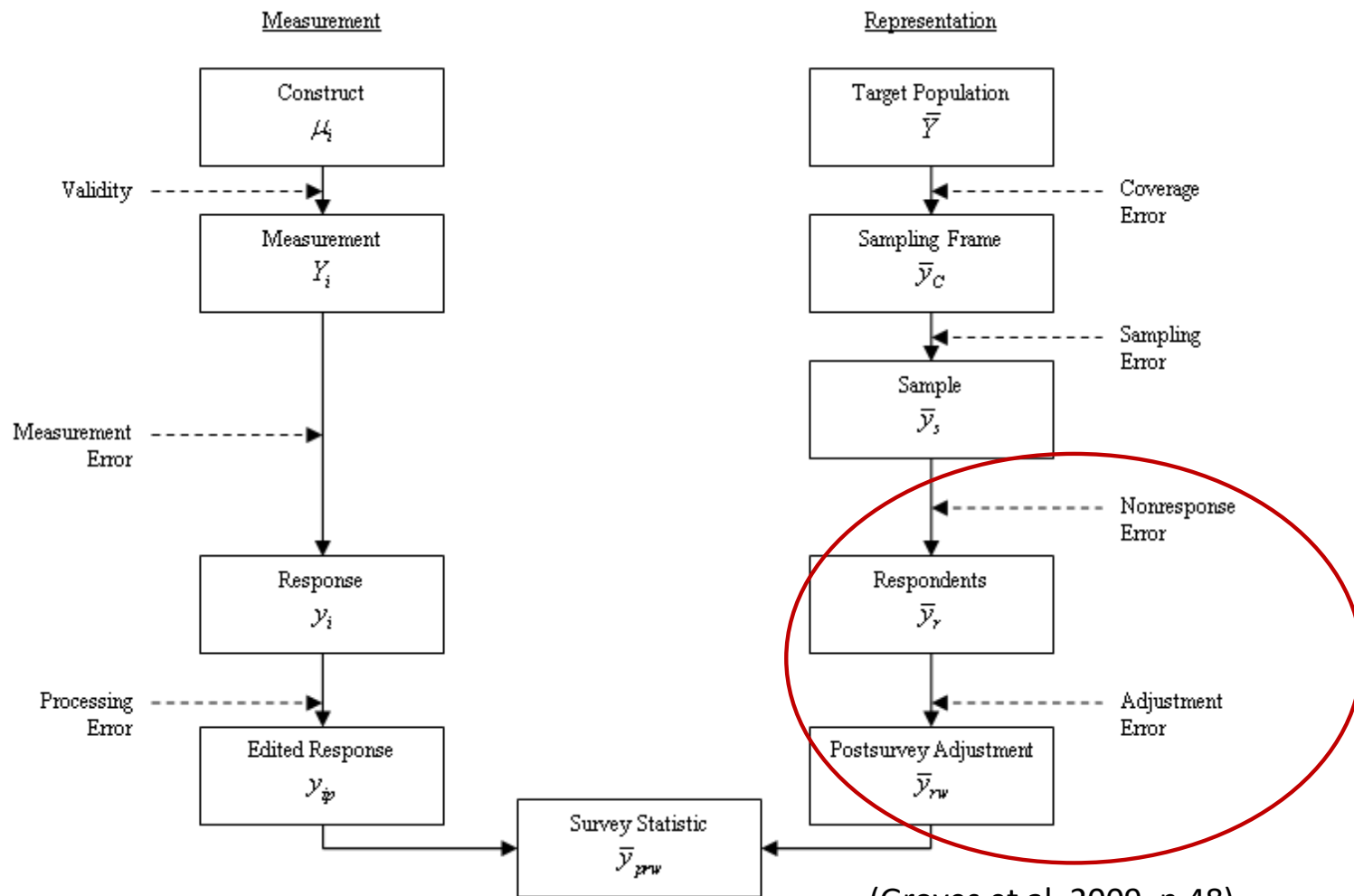
# Today

- Lecture on NR
- Exercise working with weights

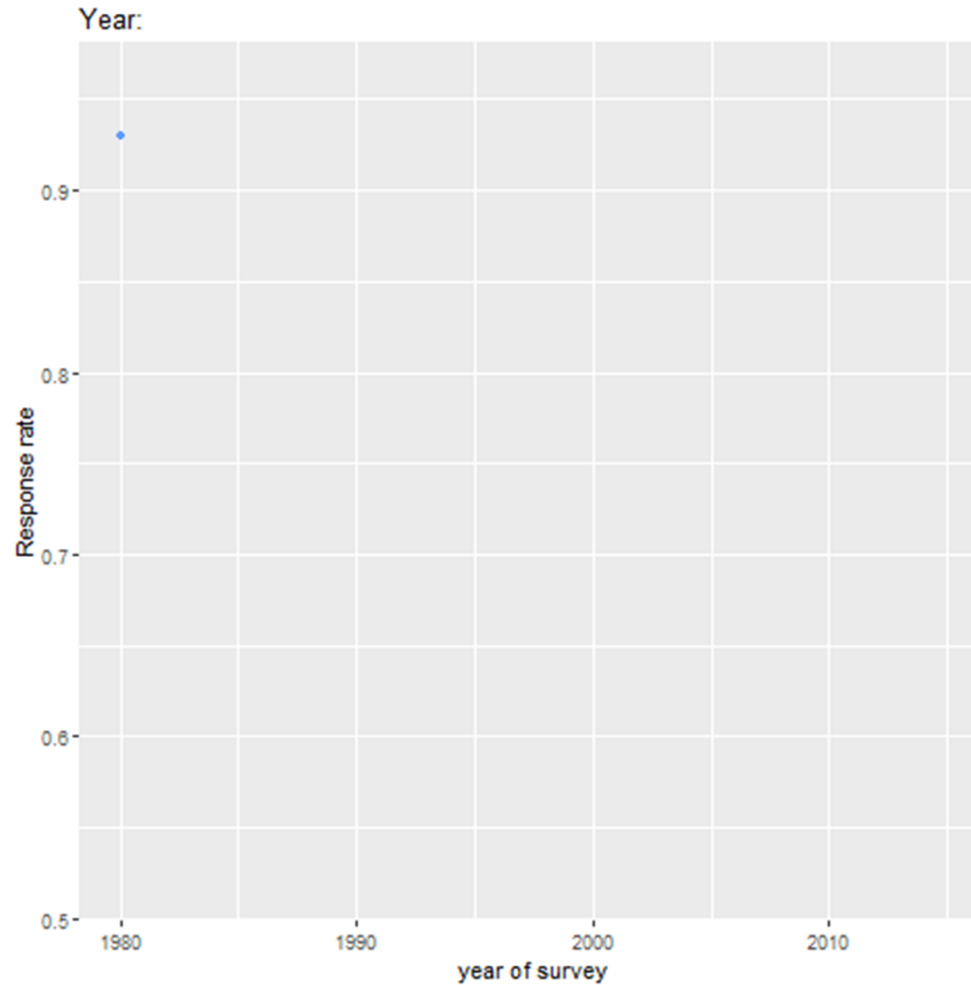
# Literature today

- The increasing problem of nonresponse
- Details of weighting methods
  - Kalton and Flores-cervantes (2003)
  - Brick (2013)

# Total Survey Error (TSE) Framework

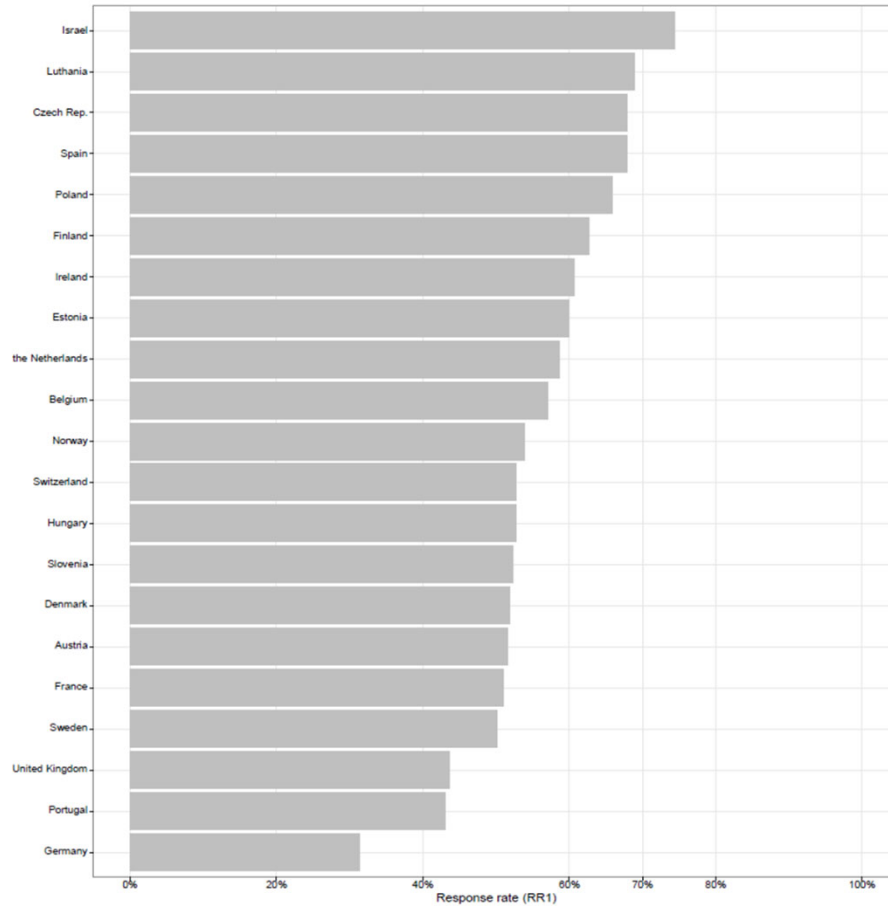


# Nonresponse in LFS over time



Based on Luiten, De leeuw & Hox (2018) International Nonresponse Trends across Countries and Years: An analysis of 36 years of Labour Force Survey data. Survey Insights: Methods from the Field. Retrieved from <https://surveyinsights.org/?p=10452>.

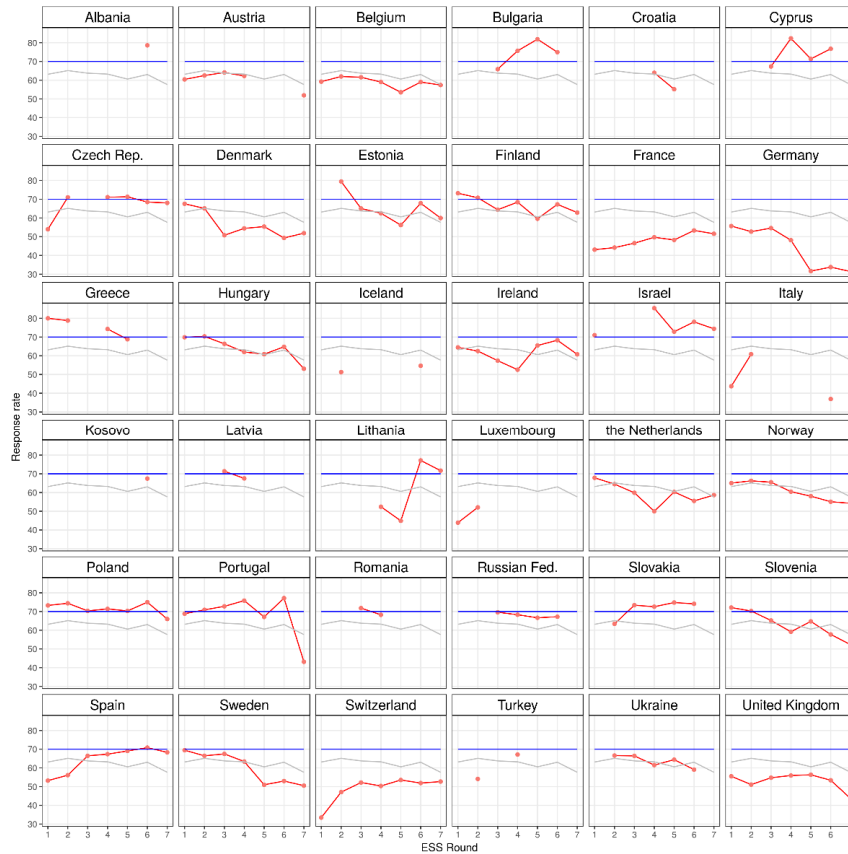
# European Social Survey Nonresponse



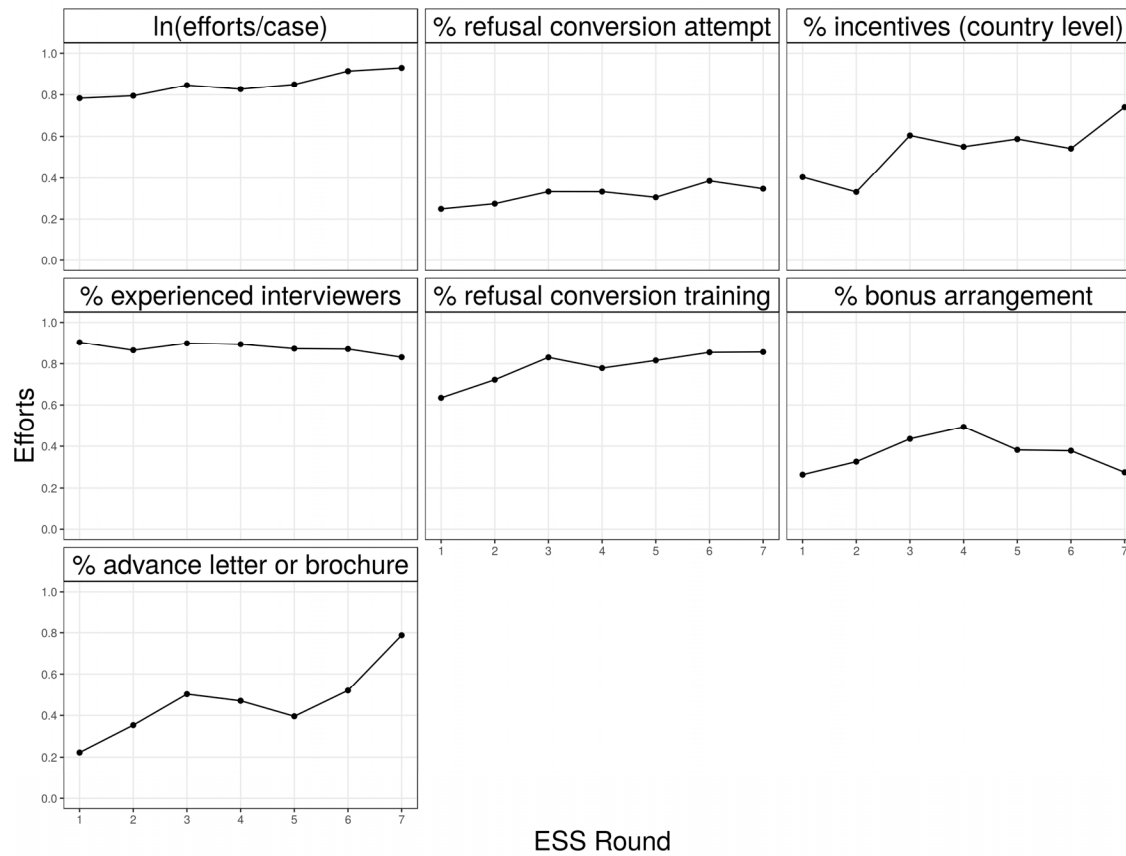
From: Beullens, K., Loosveldt G.,  
Vandenplas C. & Stoop I.  
(2018). Response Rates in the  
European Social Survey: Increasing,  
Decreasing, or a Matter of Fieldwork  
Efforts? *Survey Methods: Insights from  
the Field*. Retrieved from  
<https://surveyinsights.org/?p=9673>

**Figure 1:** Response rates per country, ESS7<sup>1</sup>

# ESS: RR variation in response rate



# ESS – increase in fieldwork efforts





# Four main types of nonresponse

In survey research we typically distinguish four types of nonresponse:

- **Unit nonresponse**  
The sample unit (e.g. person, household, institution) was sampled, i.e. belonged to the gross sample, but did not participate in the survey.
- **Item nonresponse**  
The sample unit was sampled and interviewed, but failed to provide answers to all of the survey items.
- **Attrition**  
The sample unit was sampled and initially interviewed for a longitudinal surveys, but did not complete all waves of the survey.
- **Partial (household) nonresponse**  
The sample unit was sampled and at least one member of the unit interviewed. However, at least one member of the unit did not participate.

# Main causes of nonresponse

- **Unit nonresponse**  
non-contact, refusal, unable
- **Item nonresponse**  
refusal, don't know, breakoff
- **Attrition**  
non-location, non-contact, refusal, unable
- **Partial (household) nonresponse**  
non-contact, refusal, unable

# How to prevent nonresponse

- Things you noticed in your adopted survey?

# How to prevent nonresponse

1. A good questionnaire, invitation letter, etc.
  - keep it simple, keep it simple, test it
2. Incentives
  - Preferable unconditional, and cash
3. Multiple contact attempts
4. Multiple modes (e-mail, mail, phone, f2f)
5. Refusal conversion
  - Interviewer training
6. Be responsive to questions/remarks/problems

# Correction for nonresponse

- Item nonresponse
  - Rich information on individual
- Partial (household) nonresponse
  - Proxy-answers, information on household
- Attrition
  - Information from earlier waves
- Unit nonresponse
  - Weak individual information (only frame)

Imputation

weighting



## What is nonresponse bias?

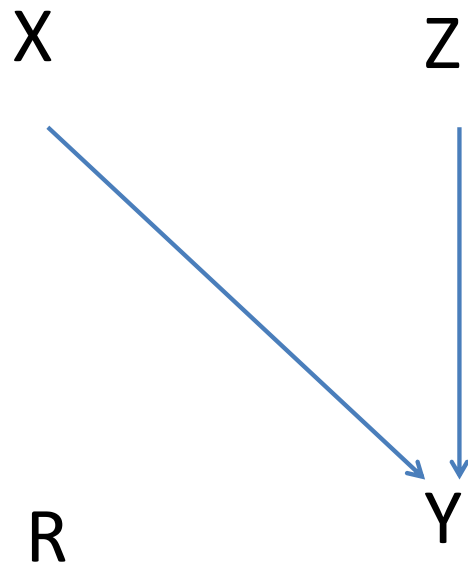
- Nonresponse bias occurs when the sampled units (e.g. individual, household, business ...) are not or only partially observed (e.g. interviewed)
- **AND** observed units are systematically different from unobserved units.

# MCAR, MAR, NMAR

- Missing Completely At Random (MCAR):  
The responding units are a random subsample of the gross sample.
- Missing At Random (MAR):  
The responding units are not a random subsample of the gross sample.  
However, the auxiliary information  $x$  renders the relationship between  $y$  and response  $r$  independent.
- Not Missing At Random (NMAR):  
The responding units are not a random subsample of the gross sample. In addition, the auxiliary information  $x$  does not render the relationship between  $y$  and response  $r$  independent.

# Missing data mechanisms

MCAR

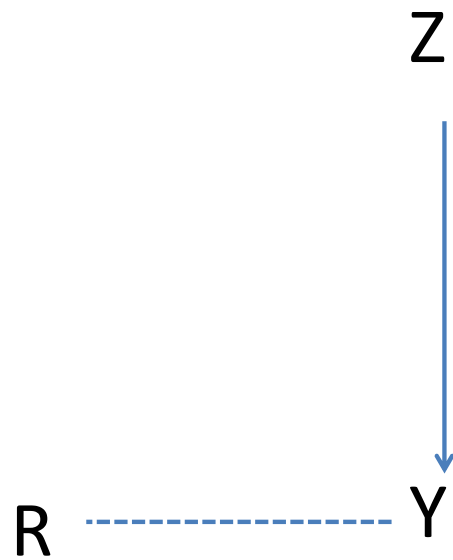




# Missing data mechanisms

MCAR

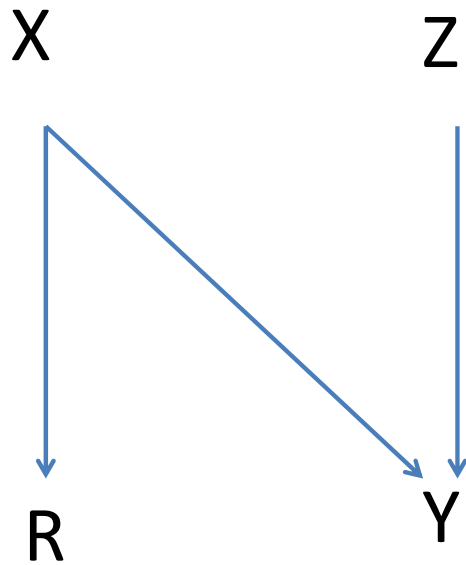
MAR - before



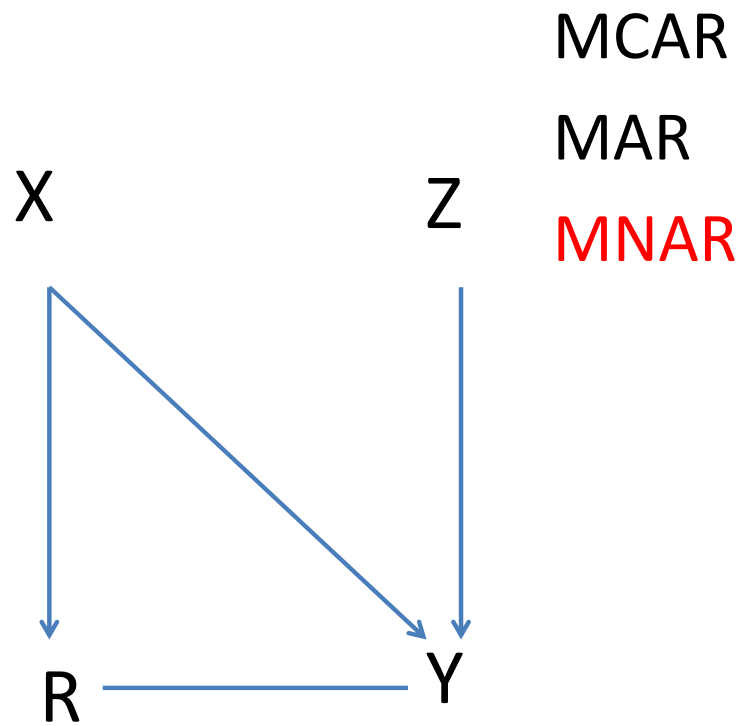
# Missing data mechanisms

MCAR

MAR - after



# Missing data mechanisms

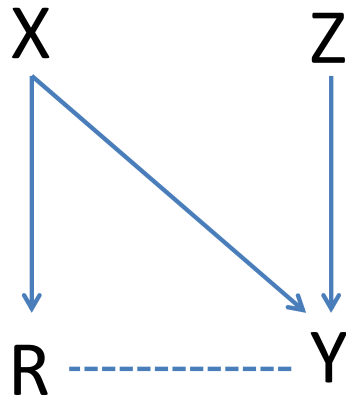


## Example: income

- Unit response rates in surveys:
  - ~5-50%
  - Nonresponse: 50-95%!
- Item-nonresponse for income question:
  - ~25%
- What do we have: MCAR, MAR, or MNAR?
  - Discuss!

# Item nonresponse (weeks 11,12)

- Use covariates (x) at level of respondent
- Strongly related to both response (R) and Y
- MCAR, MAR, MNAR models

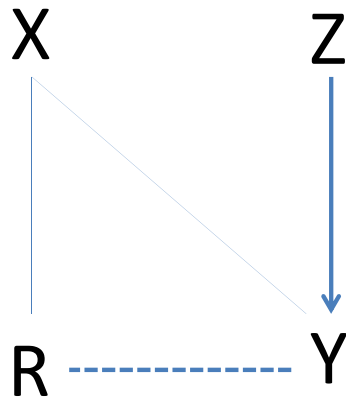


## Example item missings in income:

- Education, wealth, age, gender, value of house (X)
- To predict income (Y) and
- Take away relation R-Y

# Unit nonresponse

- Few covariates ( $x$ ) at level of respondent
  - Often only address, or e-mail
- Weakly related to both response ( $R$ ) and  $Y$



## Example Unit Nonresponse:

- Only use address (house price)
- Can predict income ( $Y$ ), but
- Cannot explain relation  $R$ - $Y$
- Not successful in NR correction

## Why weight?

**Sampling:** selection probabilities may differ

-> design weights

**Coverage:** sampling list may not cover target population

**Nonresponse:** not all people in sample will end up in data

-> adjustment weights for coverage/NR

## Design weights (repeat from weeks 3-6)

SRS: equal probabilities

- no design weights

Stratified, cluster, multistage

- need design weights for unbiased estimates See slides for those weeks
- Weights not needed if:
  - you specify correct svydesign (ids=~ , strata=~ , ....)
  - You use a HT-estimator (weight = ~inclusionprobabilities)



# Different kinds of weights

**Sampling:** selection probabilities may differ

- also known as (aka) design weights

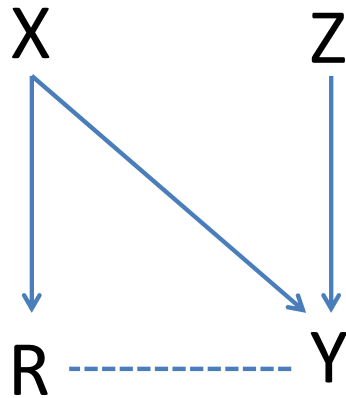
## Nonresponse weights

- **Nonresponse:** not all people in sample will end up in data
  - + **Coverage:** sampling list may not cover target population
- Aka adjustment weights

**Analysis weights = sampling weight \* nonresponse weight**

## The idea behind NR weighting

- X values at level of (sub)population
- One weighting model for all substantive analysis
  - In imputation model often Y-specific.

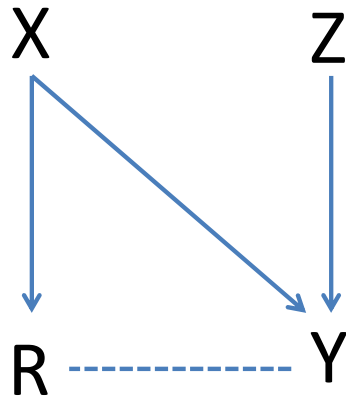


# Population level data?

1. Sampling frame (nonresponse)
  - Address.
    - Can be enriched (e.g. use google streetview)
  - Statistics Netherlands: admin data
- 2. Population level statistics (coverage + NR)
  - Gender (50/50), age, income, region, nationality
  - Next week...

## How weighting works

1. This week:  $X$  variables on sampling frame
2. Predict  $R_{0,1}$  with  $X$
3. Get predicted probabilities
4. Weight by inverse of probabilities



# Propensity-score weights

For **propensity-score weights** (logistic regression) models estimate the response propensity (**predicted probability**) of each sample unit given a set of covariates.

- Response rate for all linear combinations of for example:
  - $\text{response}[0;1] \sim \text{gender} + \text{age} + \text{region} + \text{typehouse}$
  - Save predicted probabilities

Weight is the scaled inverse of the predicted response propensity of each sample unit.

# Brick (2013)

- Review of weighting approaches
  - Propensity score models
    - Sampling frame data is limited
  - Population information is limited
    - Other weighting models
- Can we use more information?
  - Next week:
  - Section 7: paradata?
  - Population data

# Next weeks

- Next week: designing weights
- After that
  - Designed big data
  - Two weeks on missing data
  - Non-probability inference
  - Data integration
- Assignment 2 (weighting and imputation)