Microeconometrics

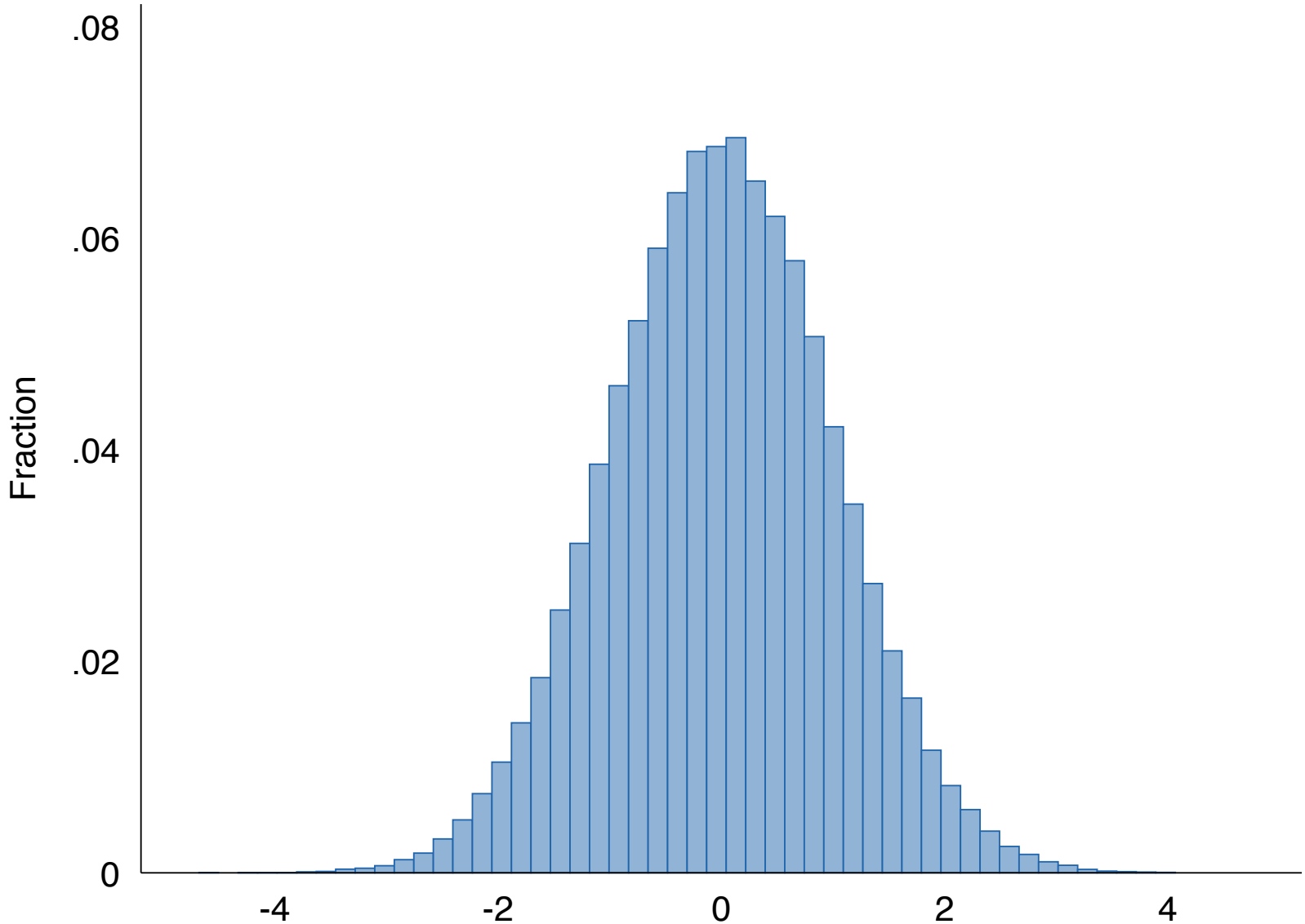
Missing data and selection models

Alex Armand
alex.armand@novasbe.pt

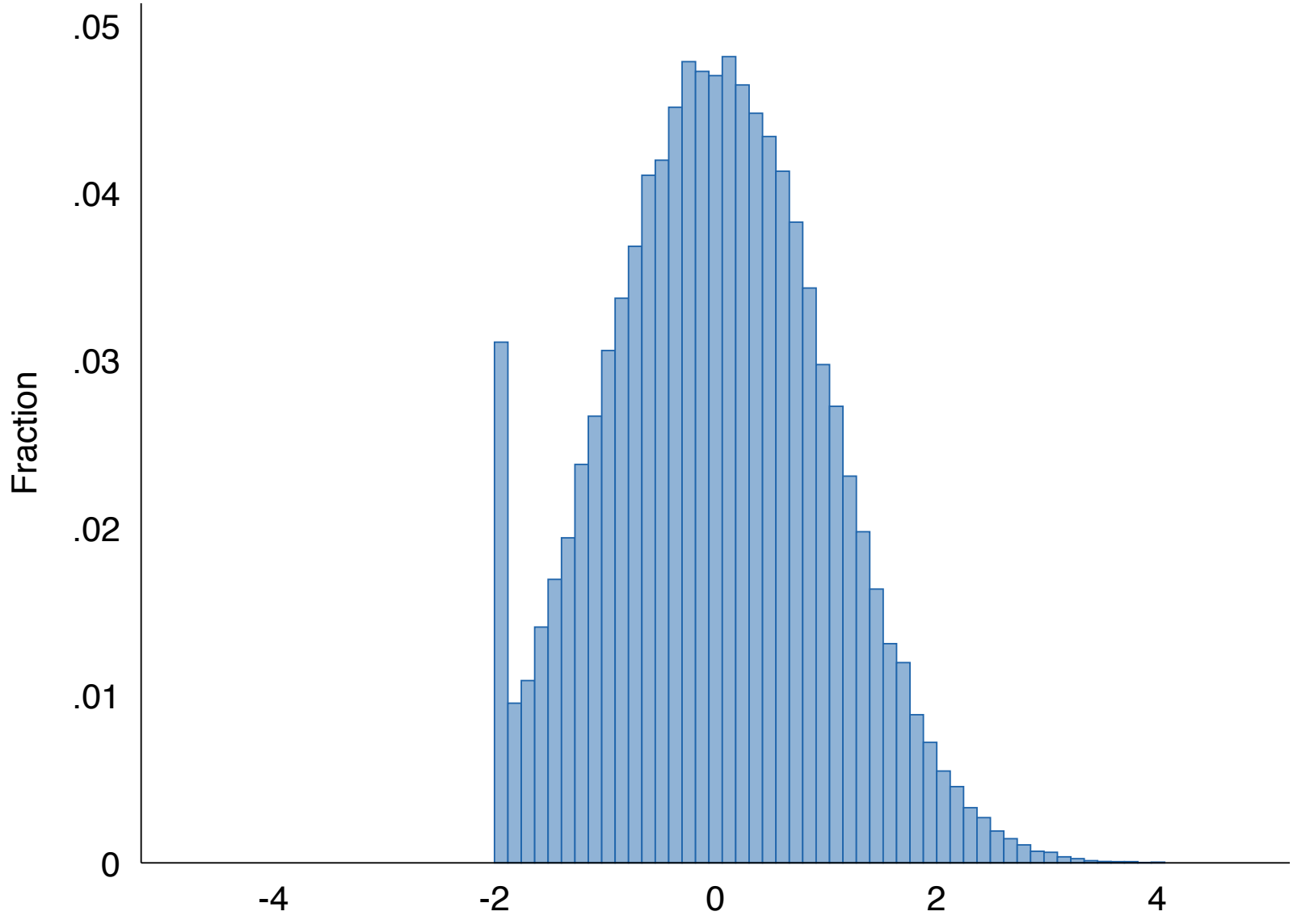
Lecture summary

1. Introduction
2. When can sample selection be ignored?
3. Selection on the response variable: truncated regression
4. Incidental truncation

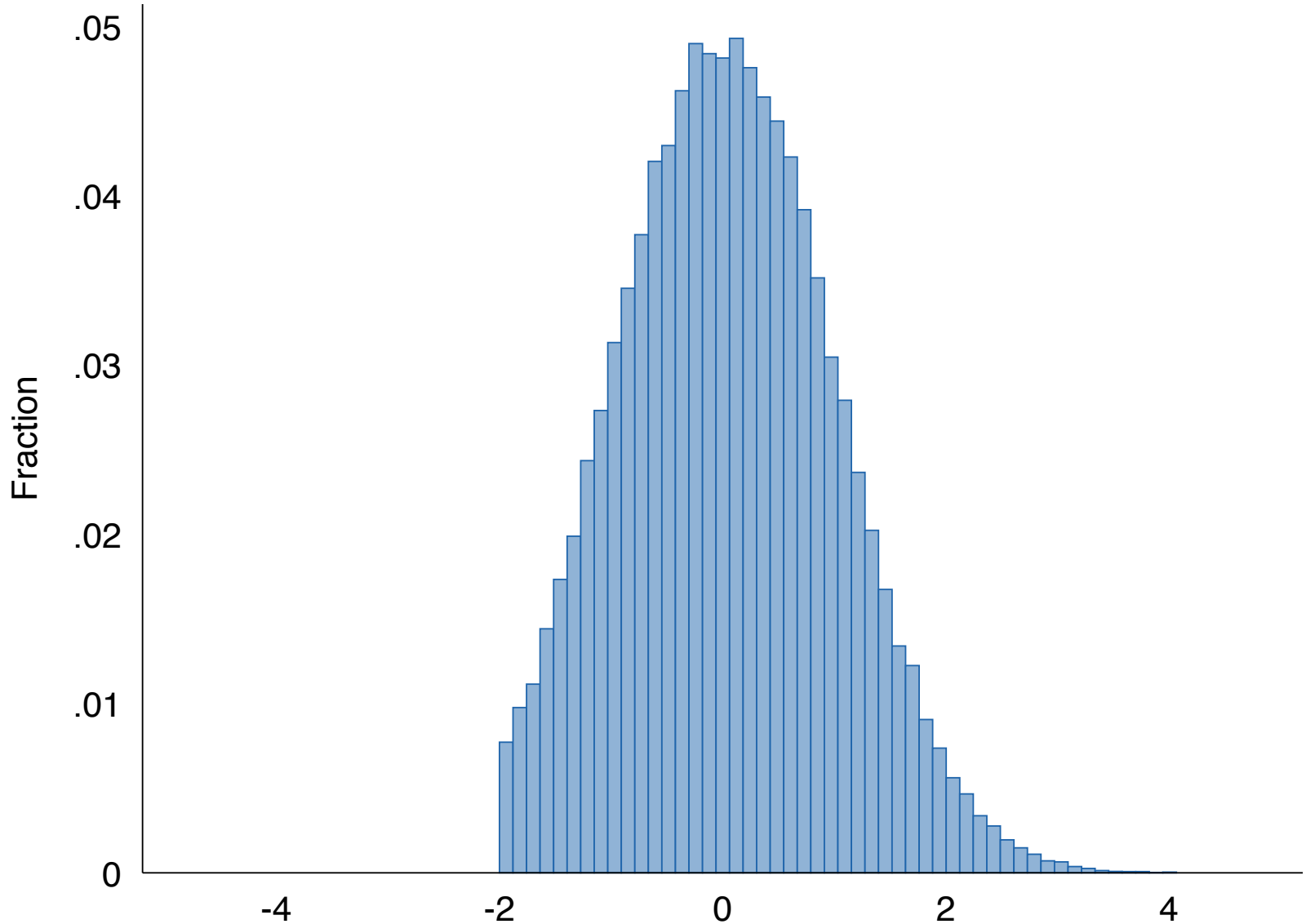
Example: Normal distribution (0,1)



Example: Normal distribution censored



Example: Normal distribution truncated



Sample selection

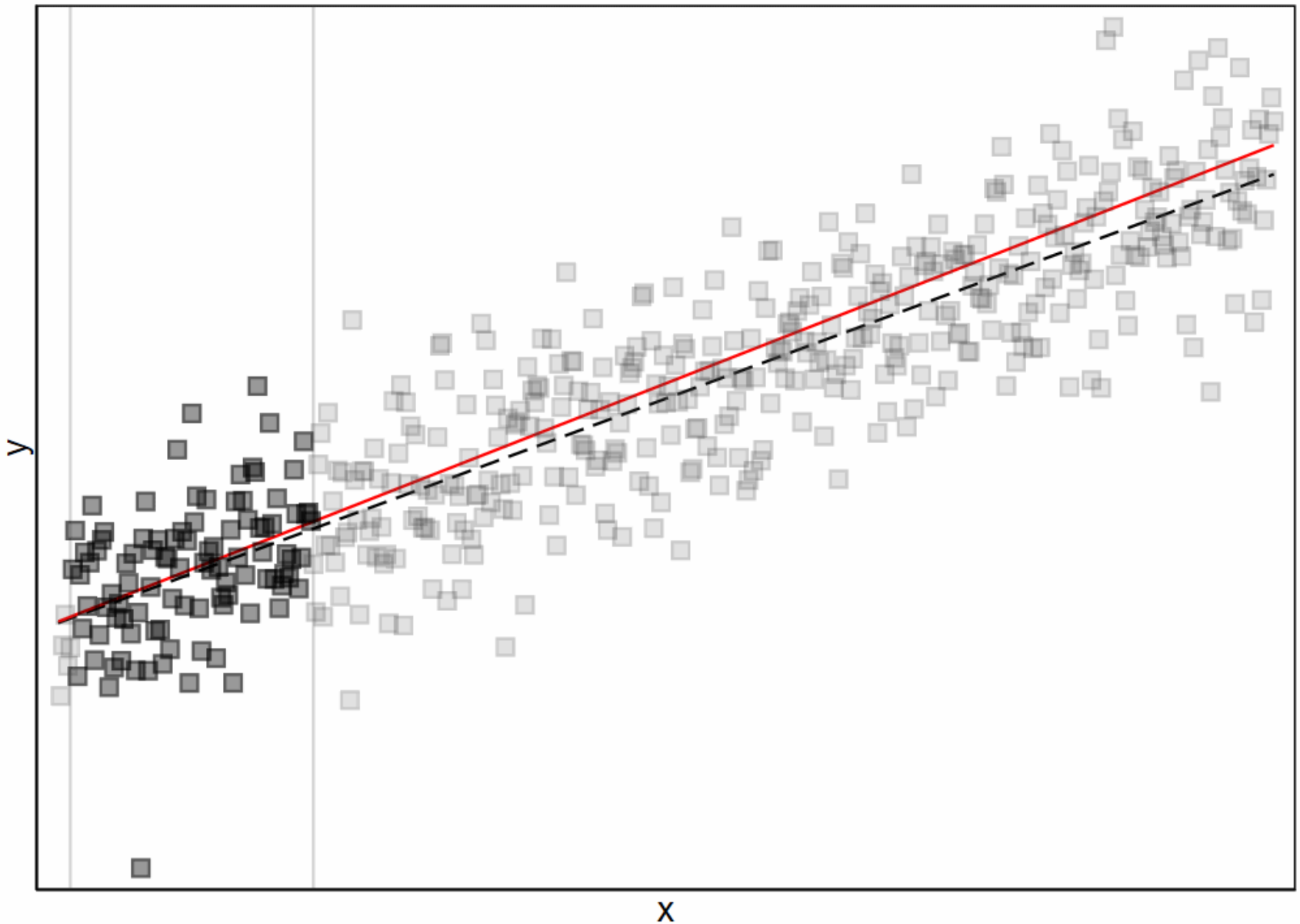
Sample selection $\Rightarrow$ the sample we obtain is not representative of the population of interest.

- **Example:** wealth equation for all families in a country

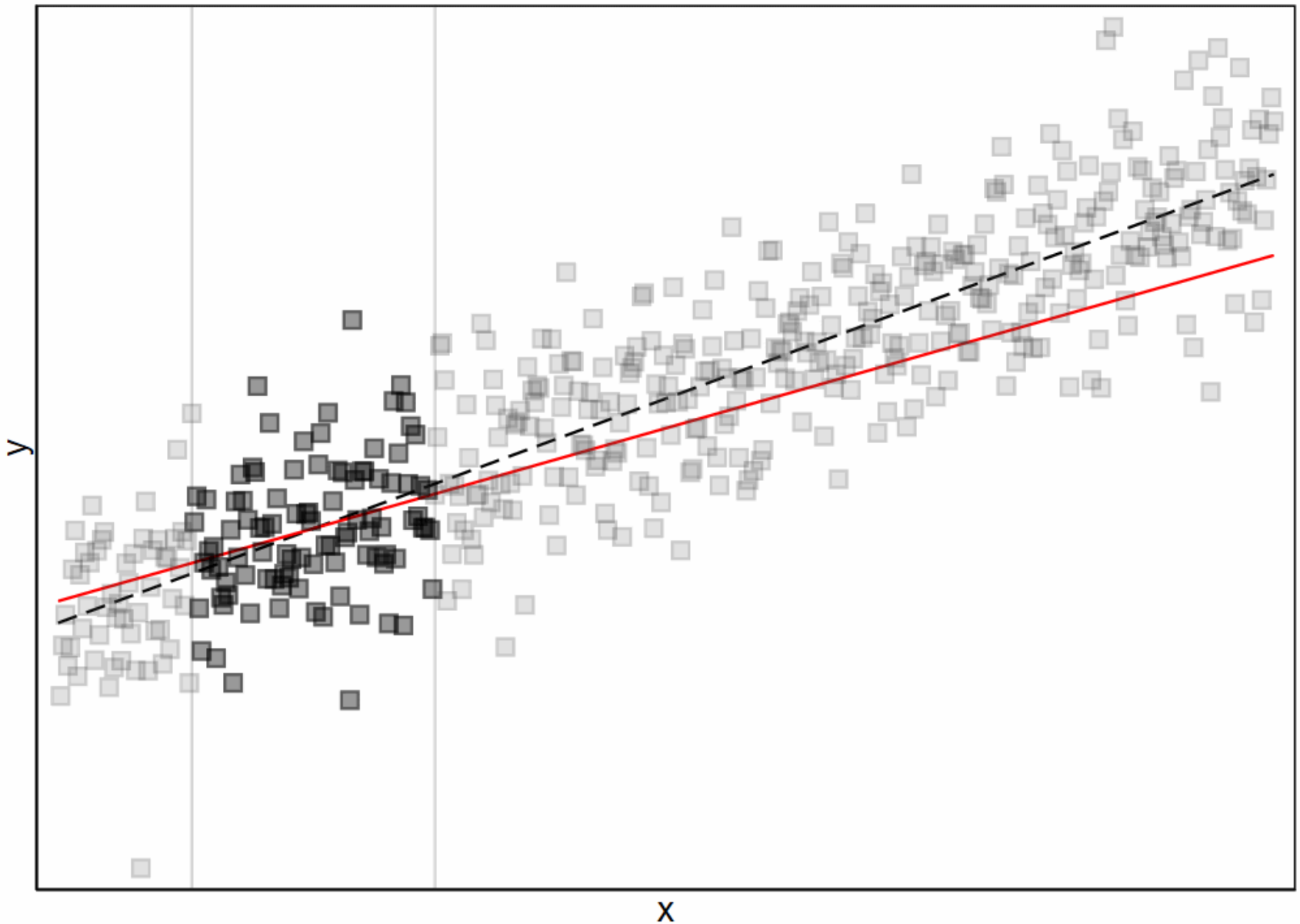
$$wealth = \beta_0 + \beta_1 plan + \beta_2 educ + \beta_3 age + \beta_4 income + u$$

1. **Random sample** $\rightarrow$ OLS
2. **Only people less than 65 years old were sampled** $\rightarrow$ selection on x
   - What if we use OLS on the **selected sample**?
3. **Only families with wealth greater than zero are sampled** $\rightarrow$ selection on y
   - What if we use OLS on the **selected sample**?

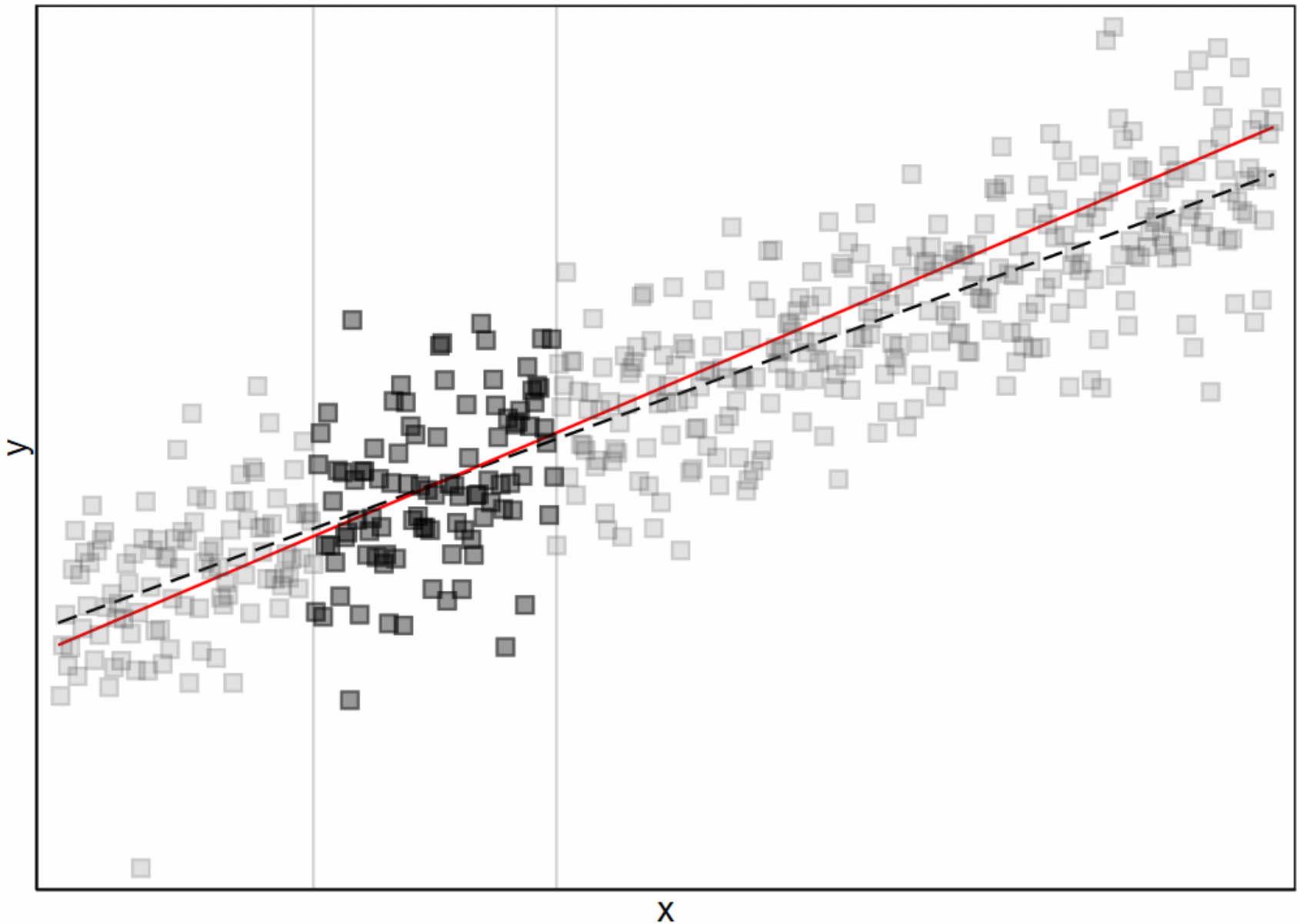
Sample selection: case 2



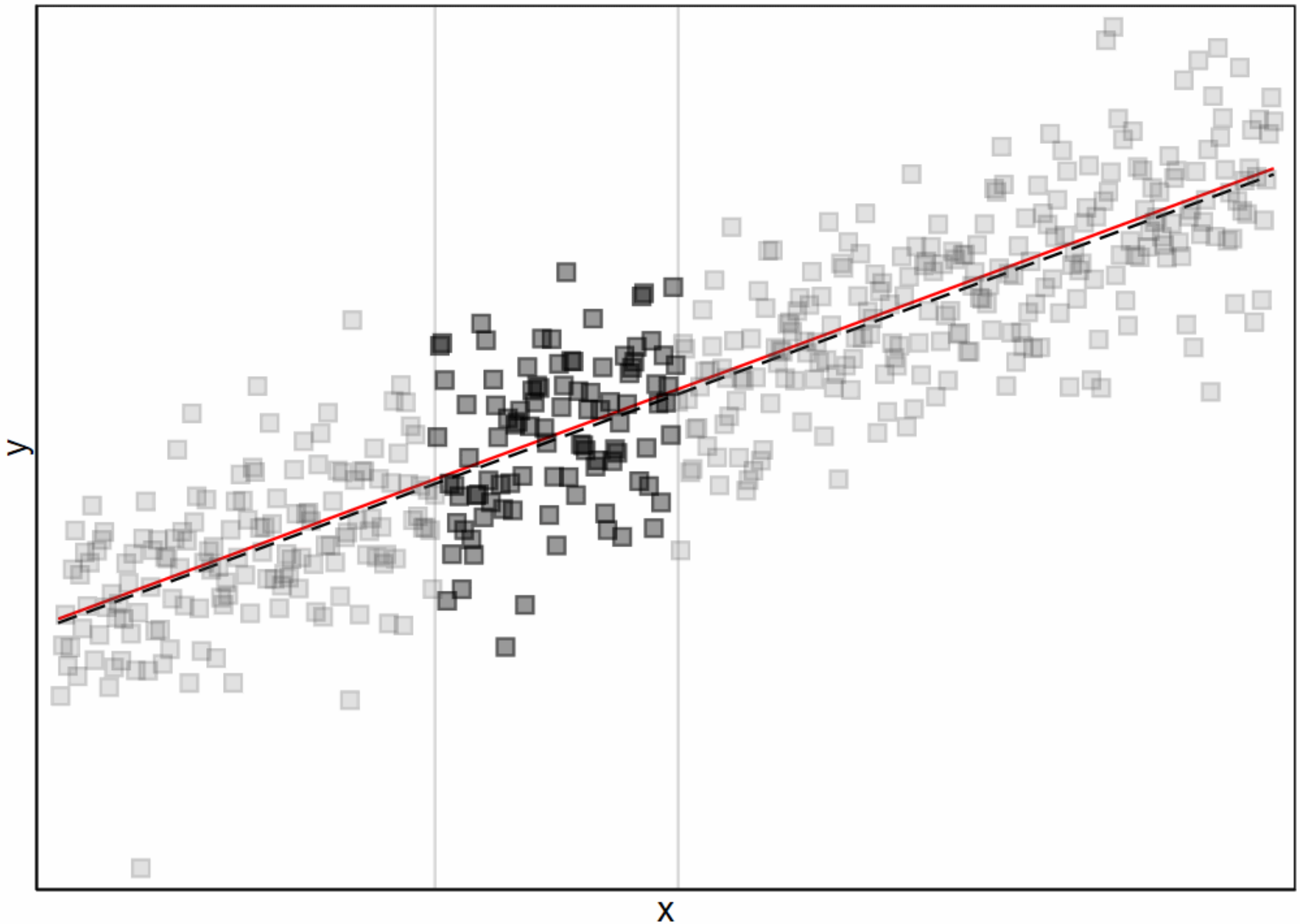
Sample selection: case 2



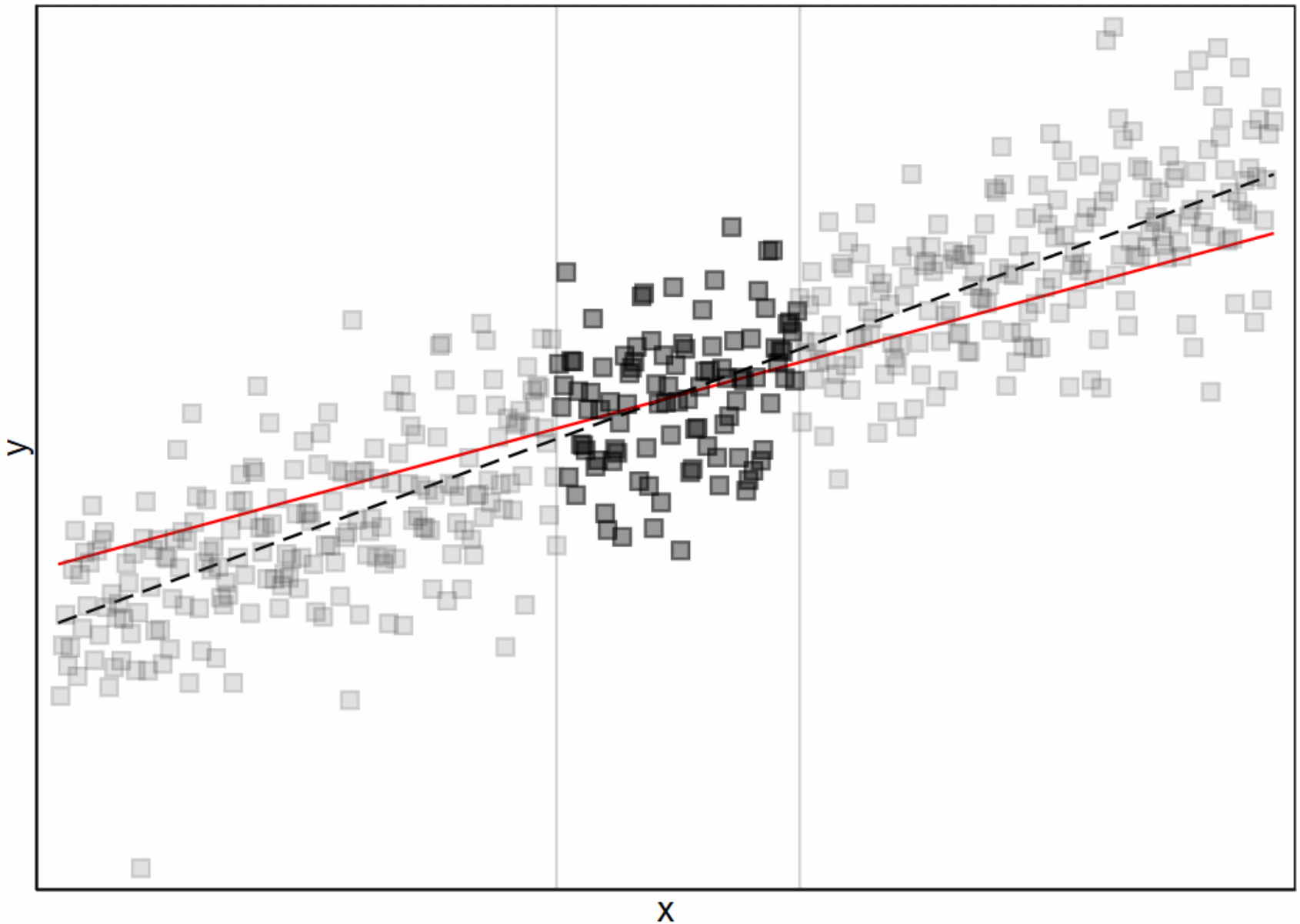
Sample selection: case 2



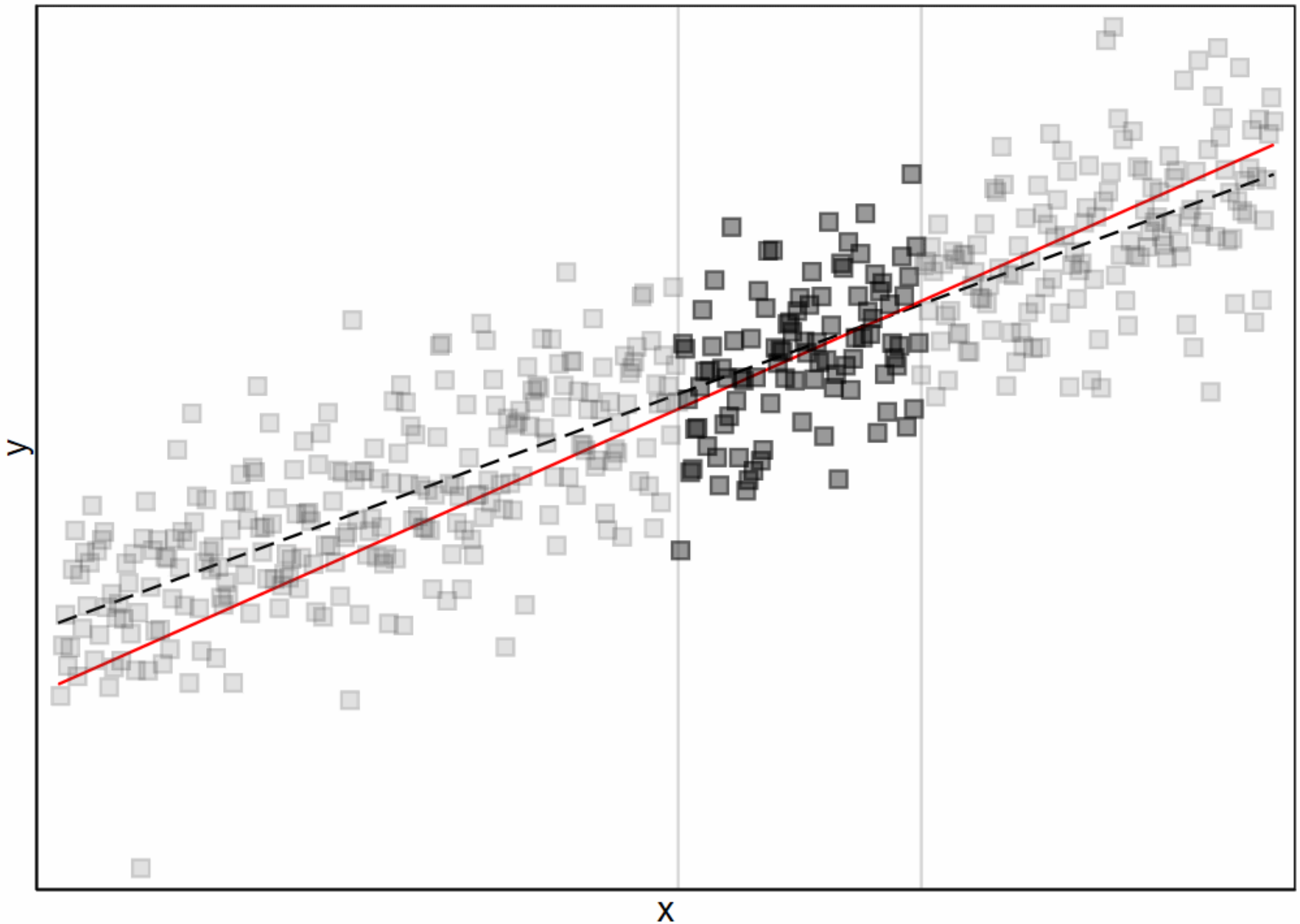
Sample selection: case 2



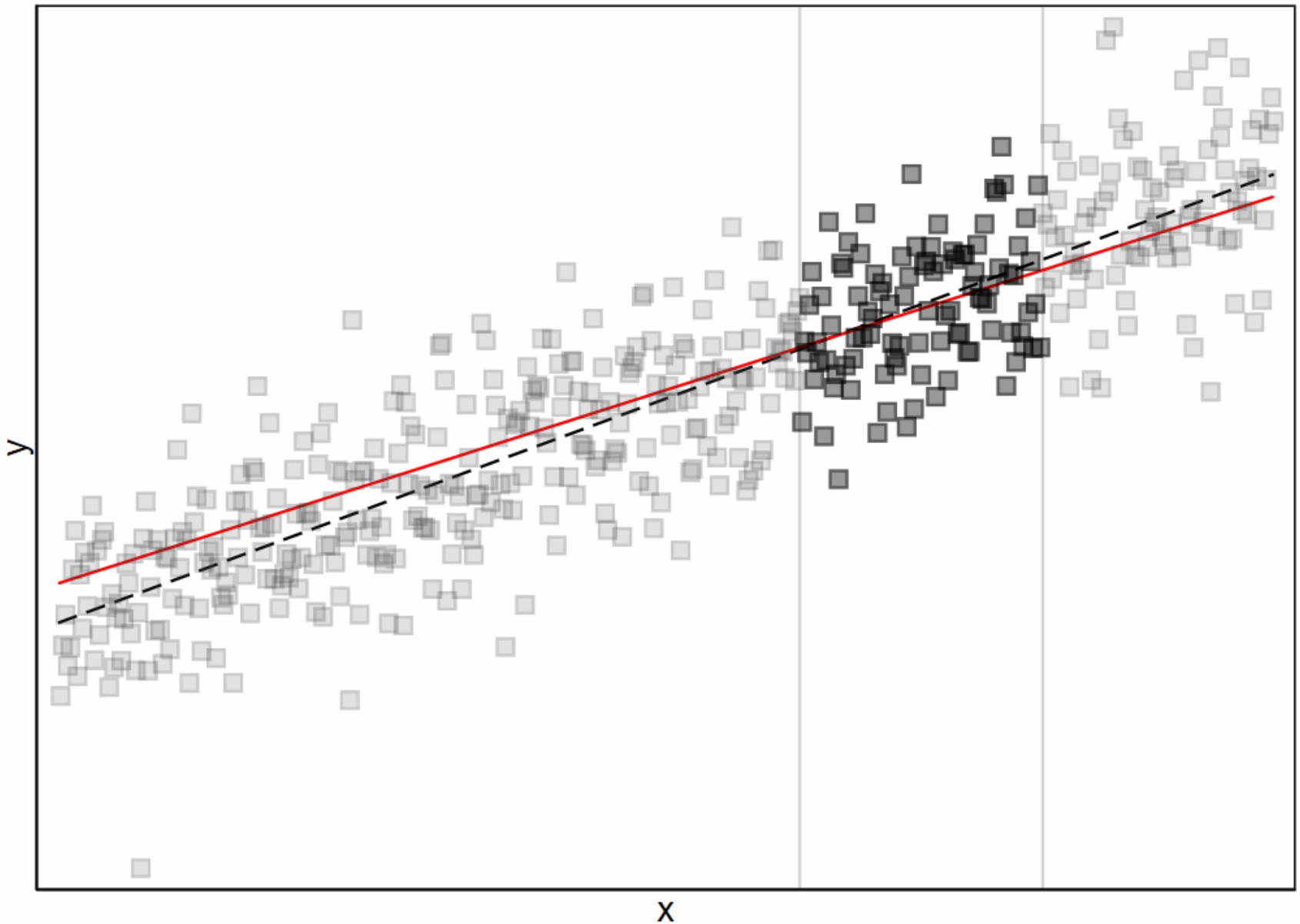
Sample selection: case 2



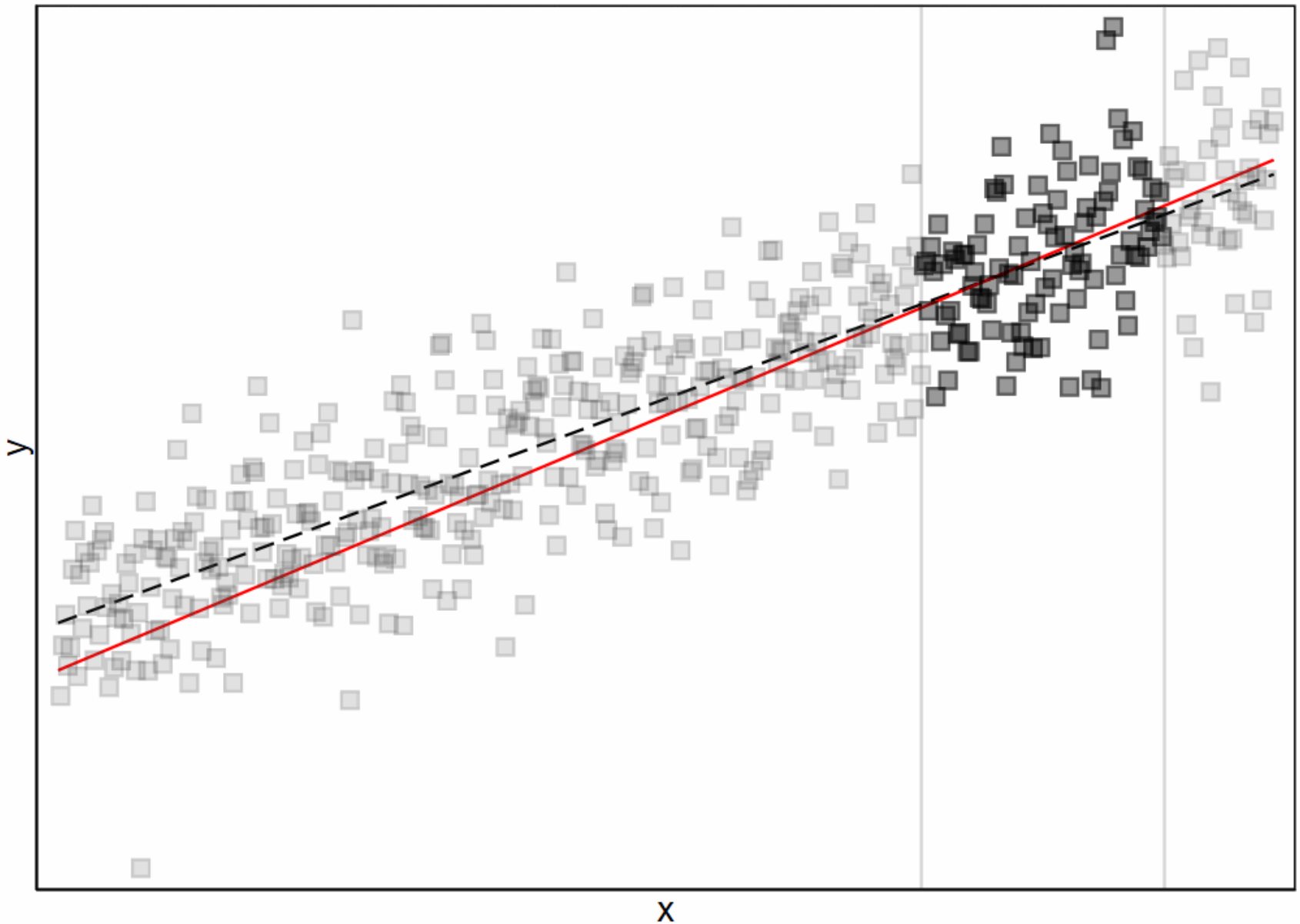
Sample selection: case 2



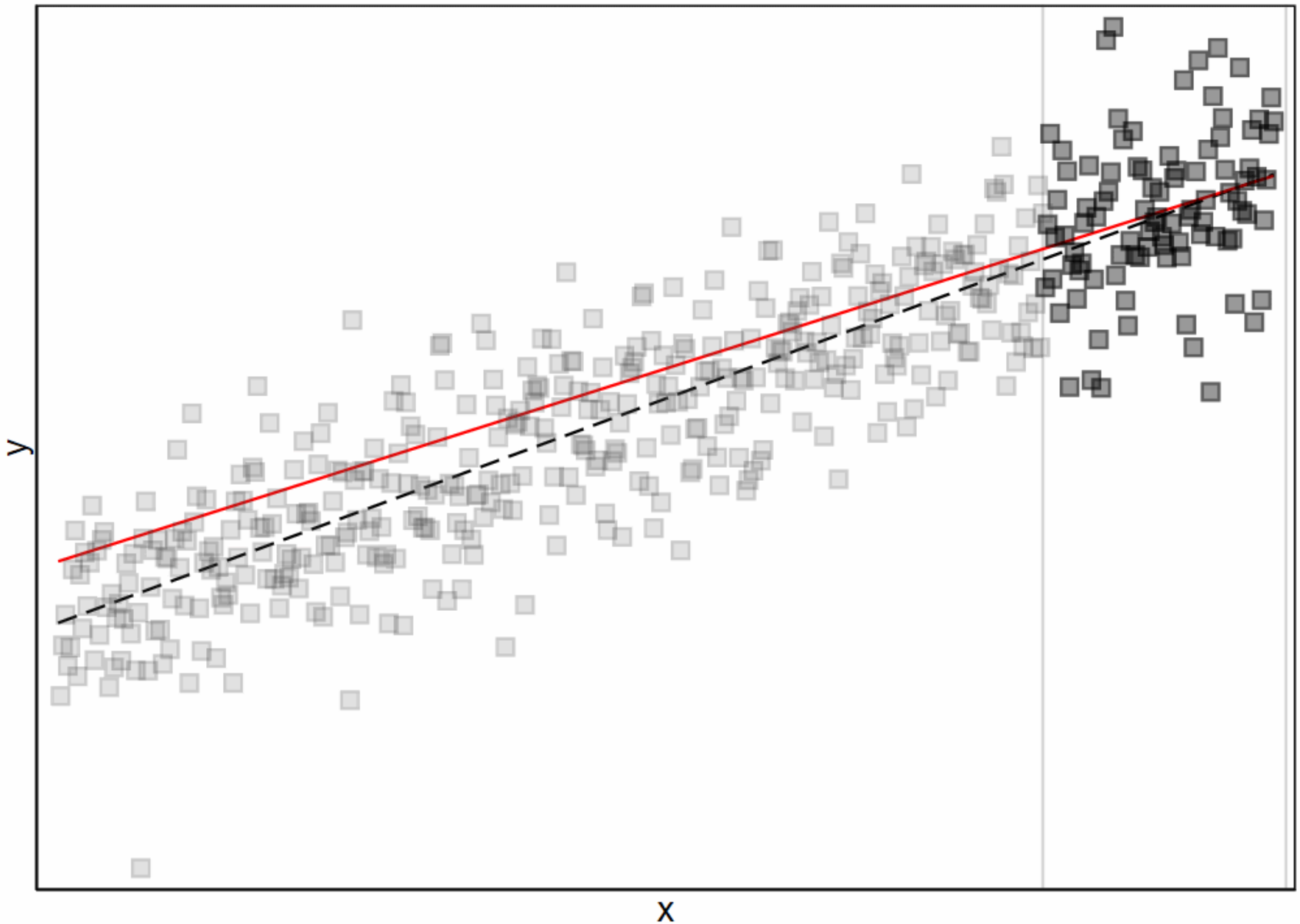
Sample selection: case 2



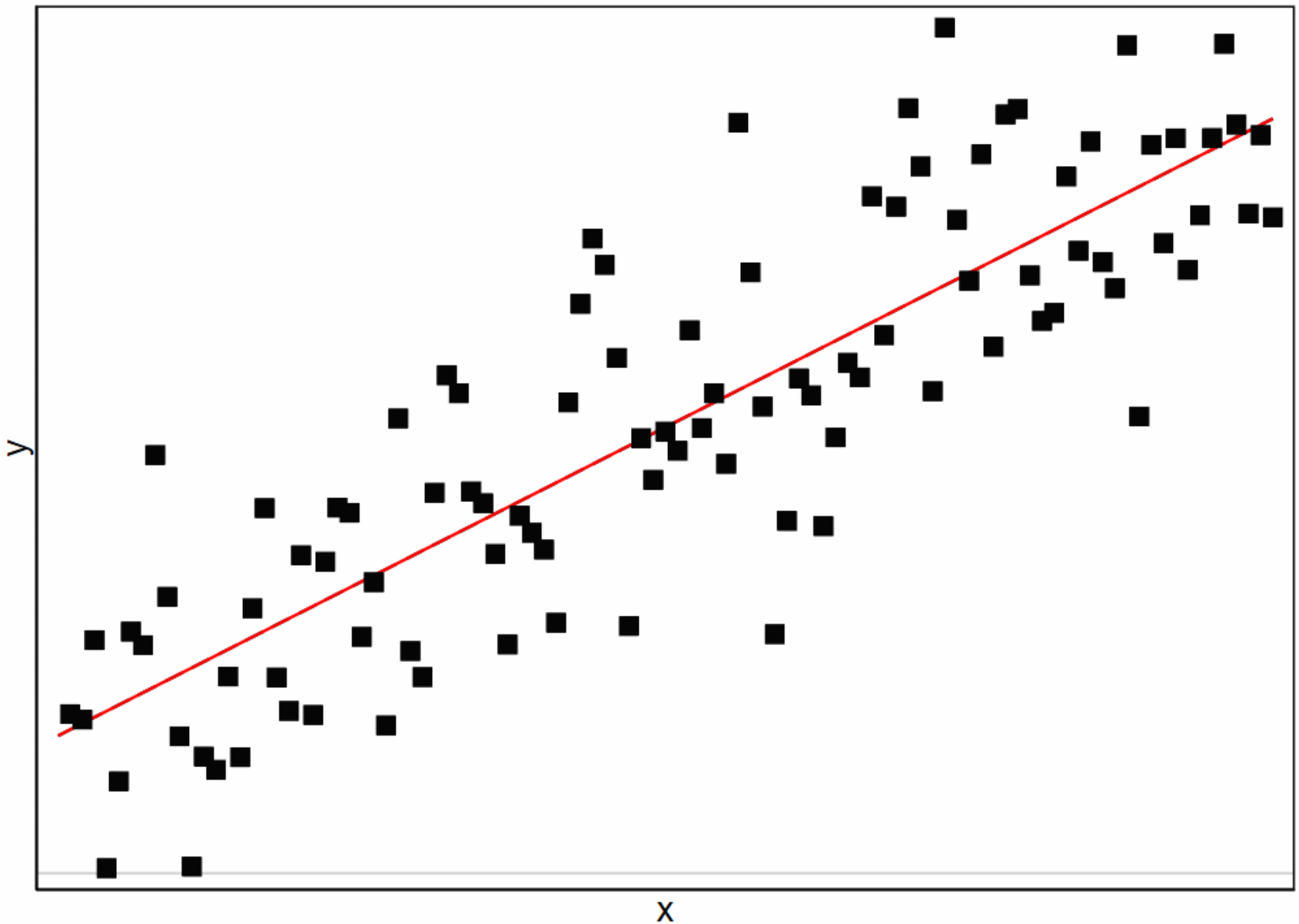
Sample selection: case 2



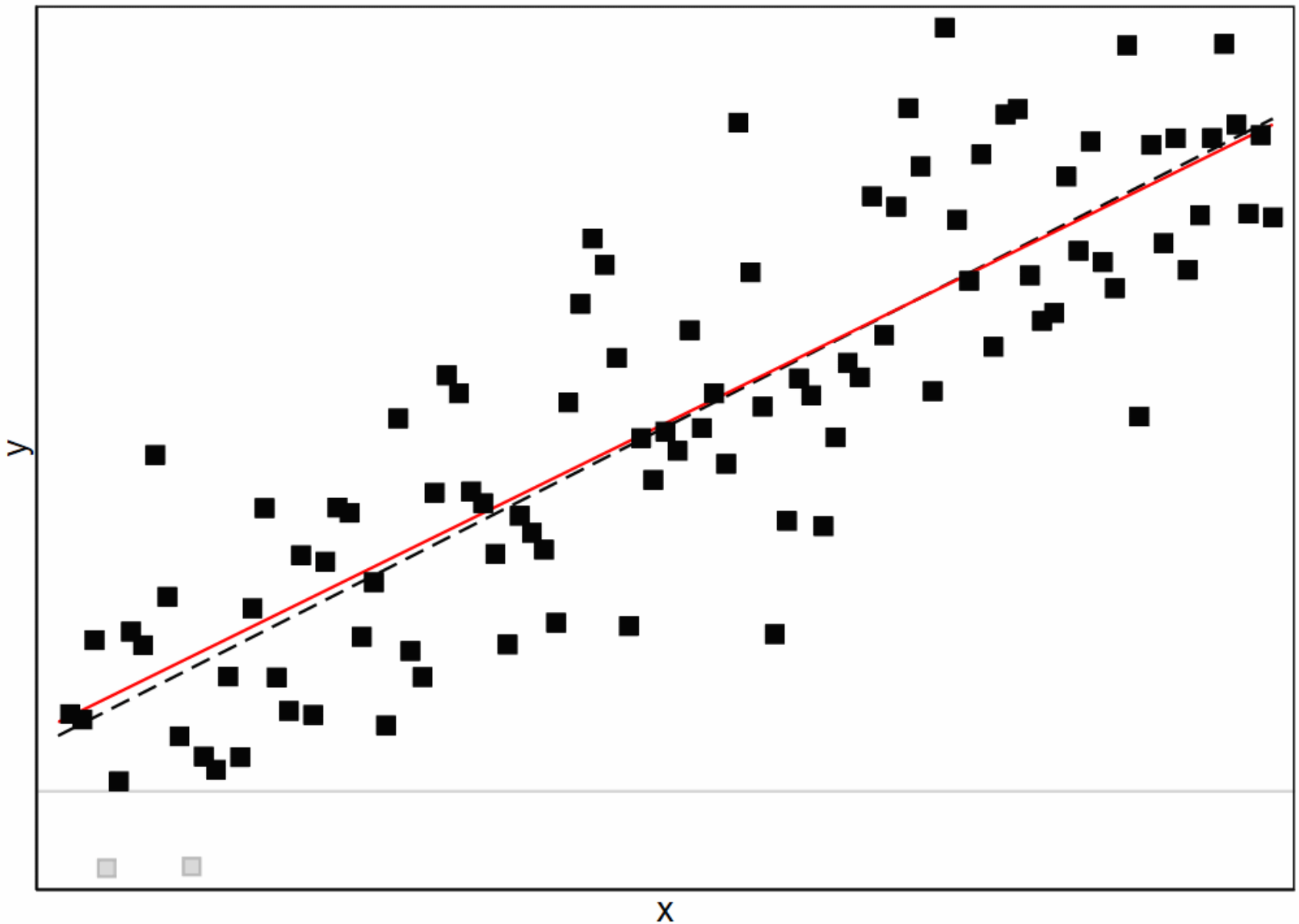
Sample selection: case 2



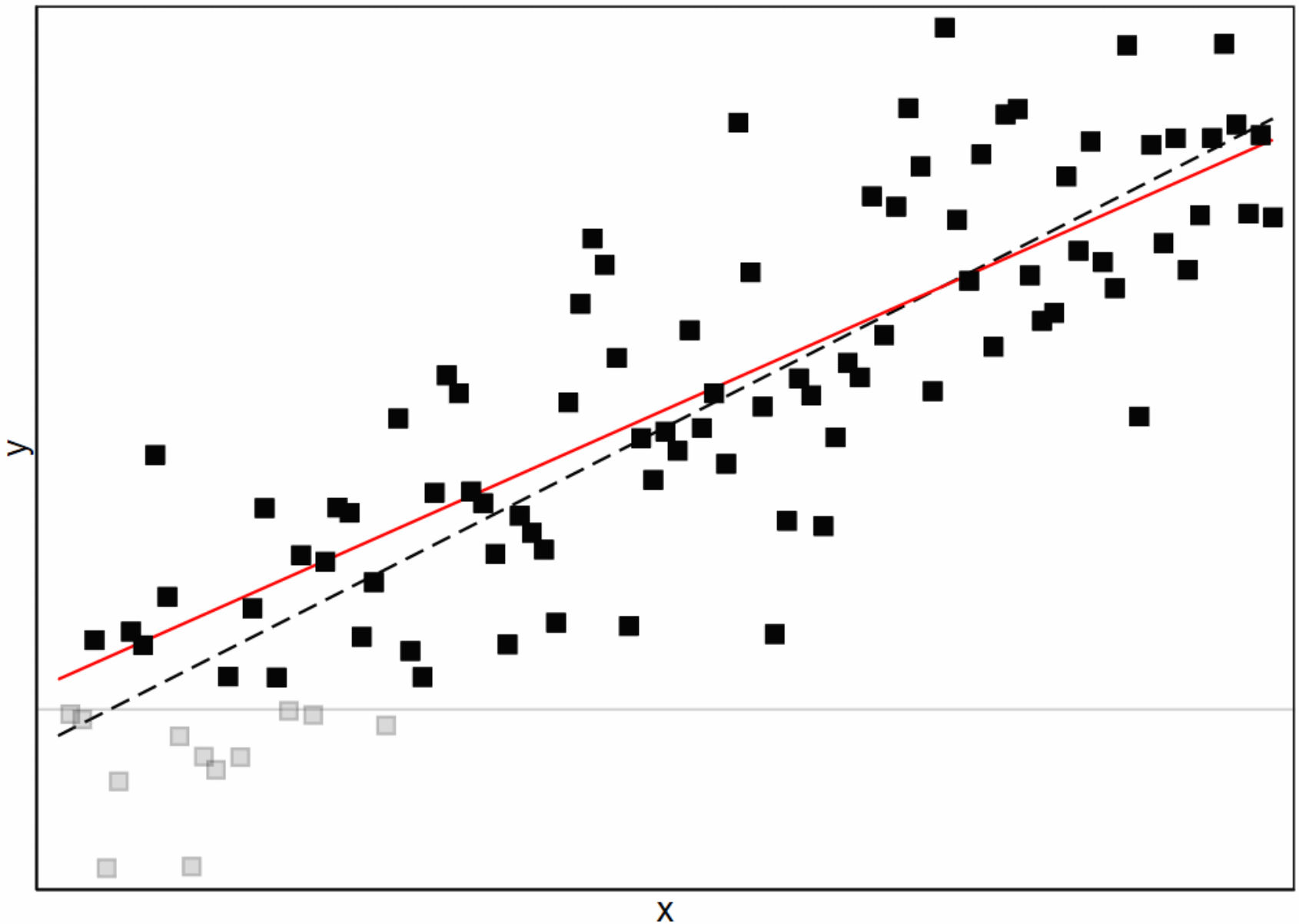
Sample selection: case 3



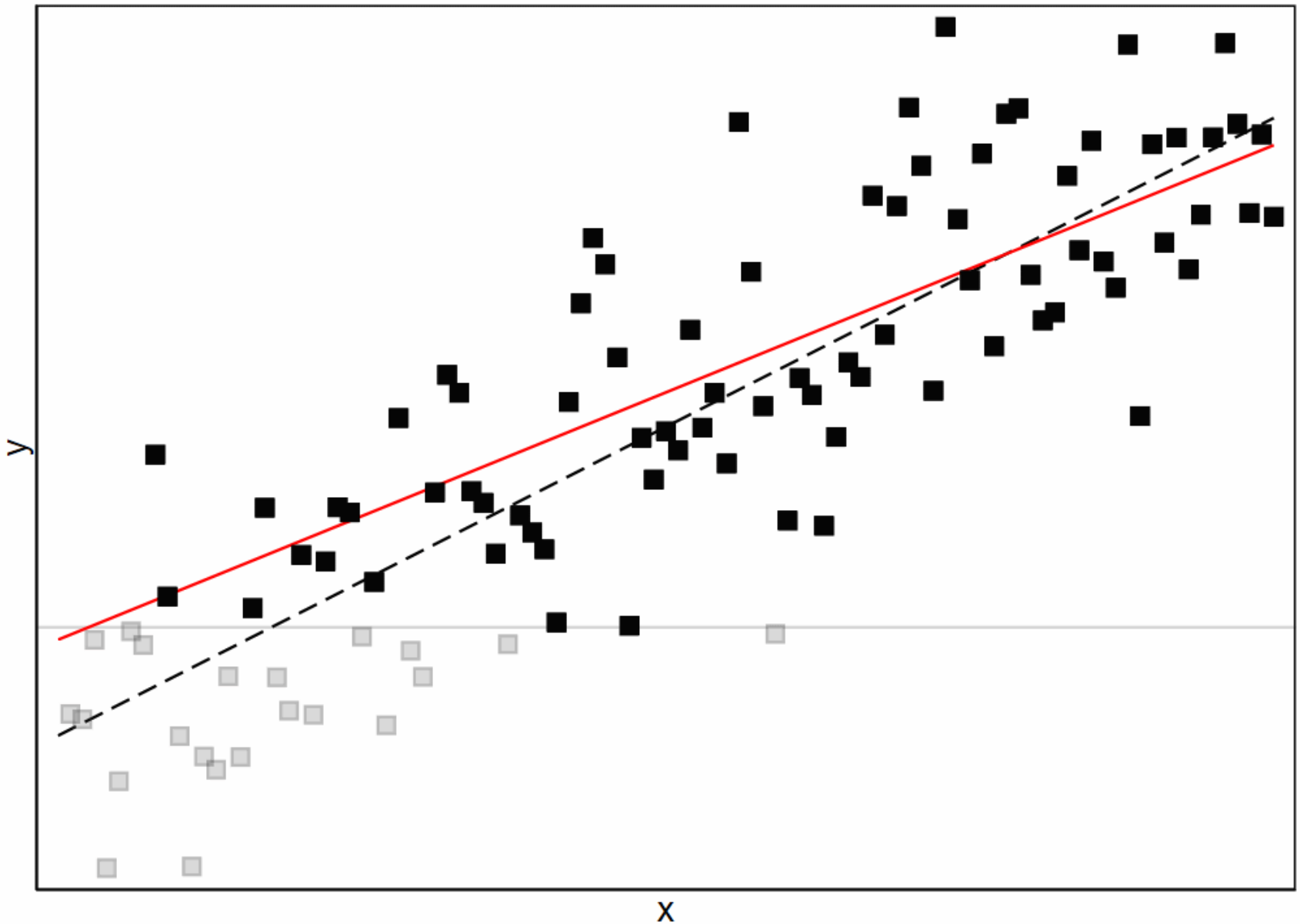
Sample selection: case 3



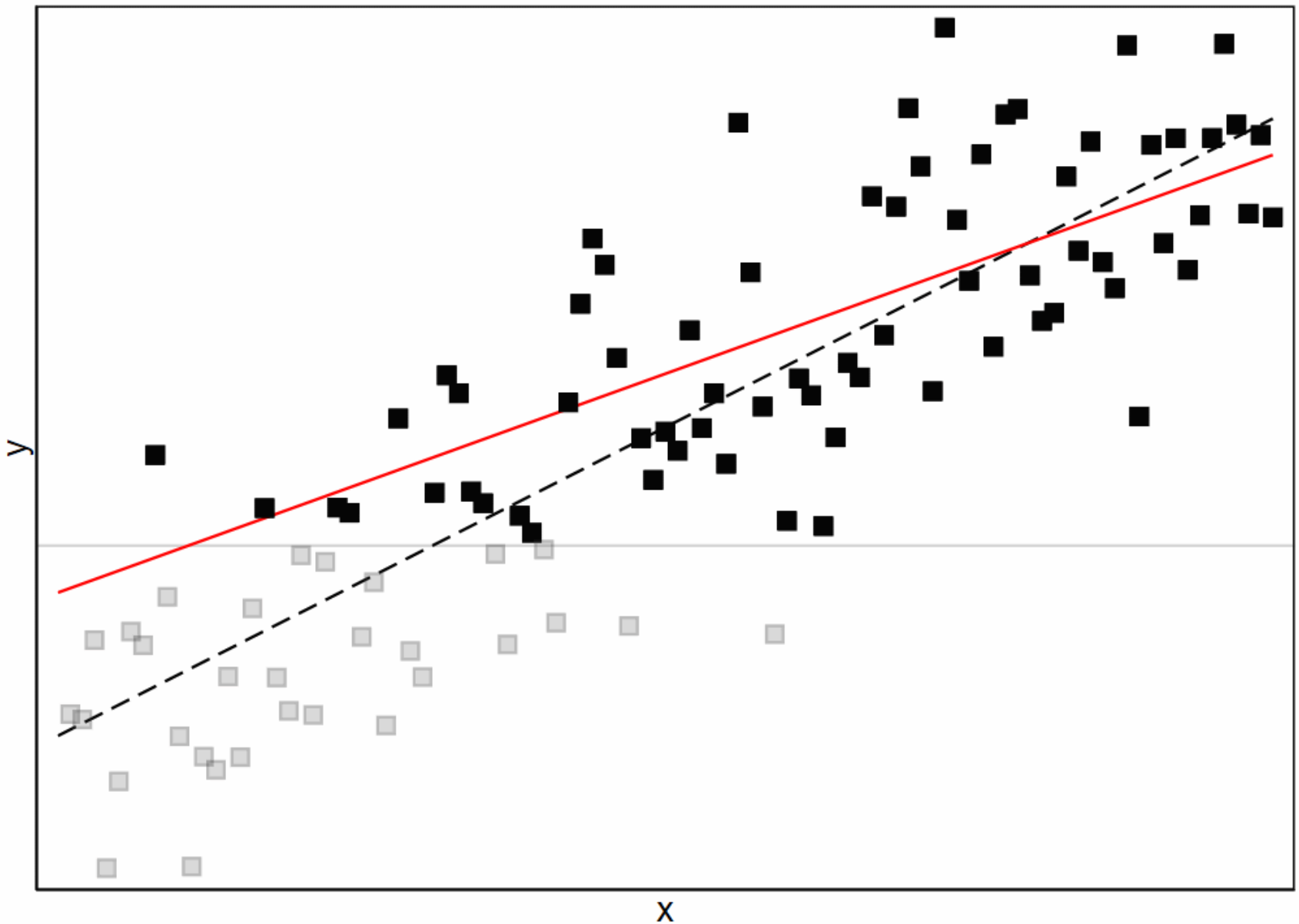
Sample selection: case 3



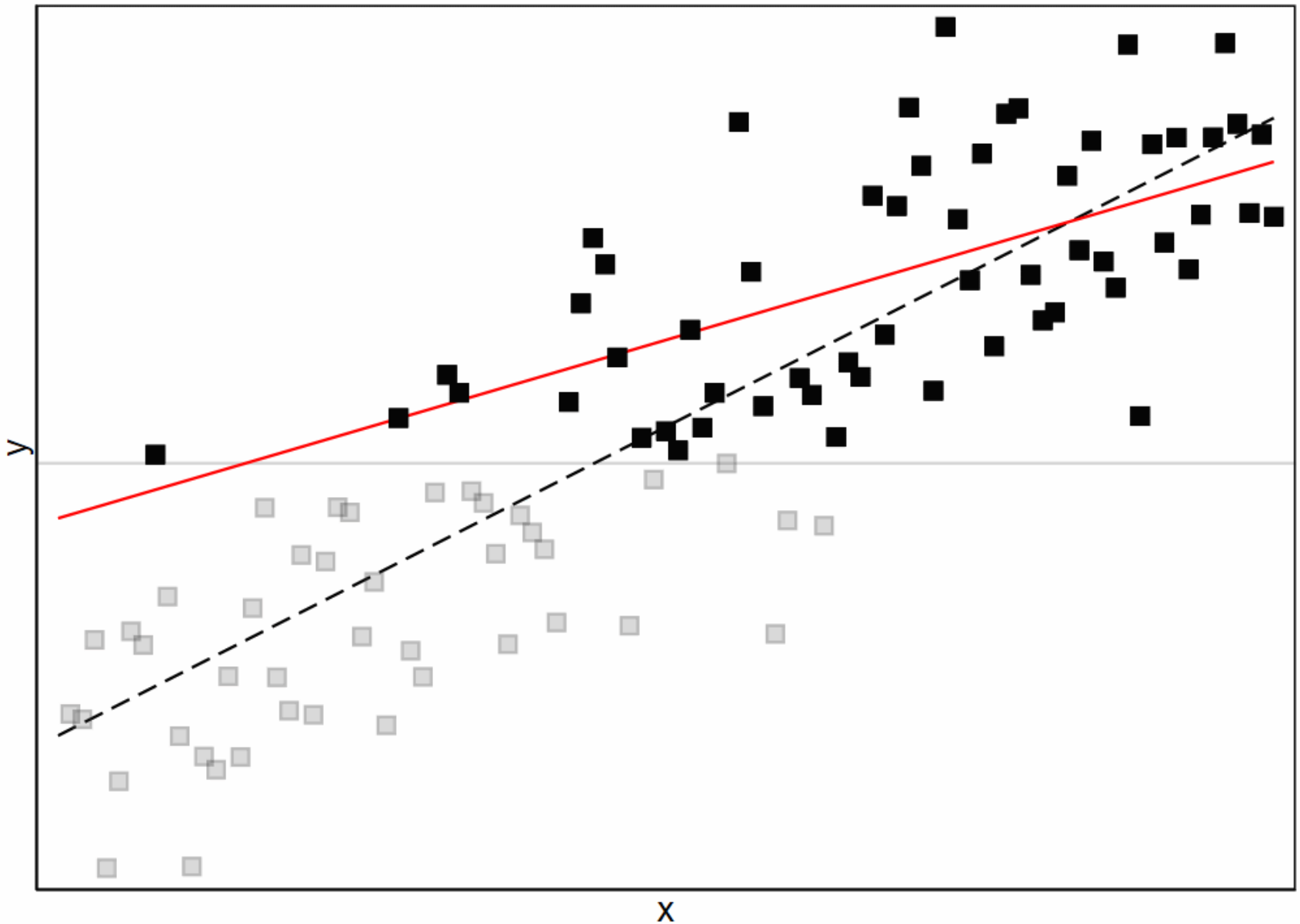
Sample selection: case 3



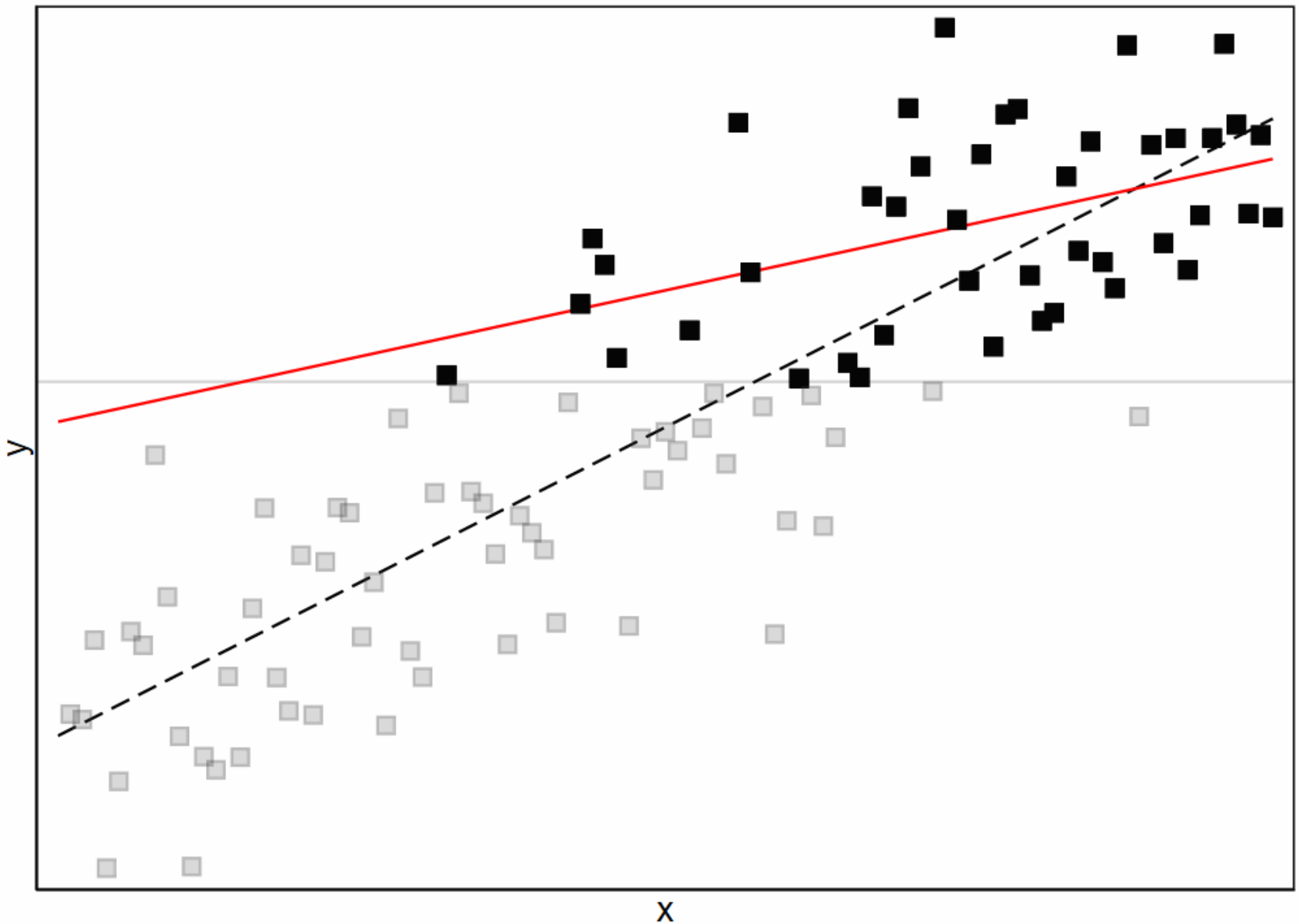
Sample selection: case 3



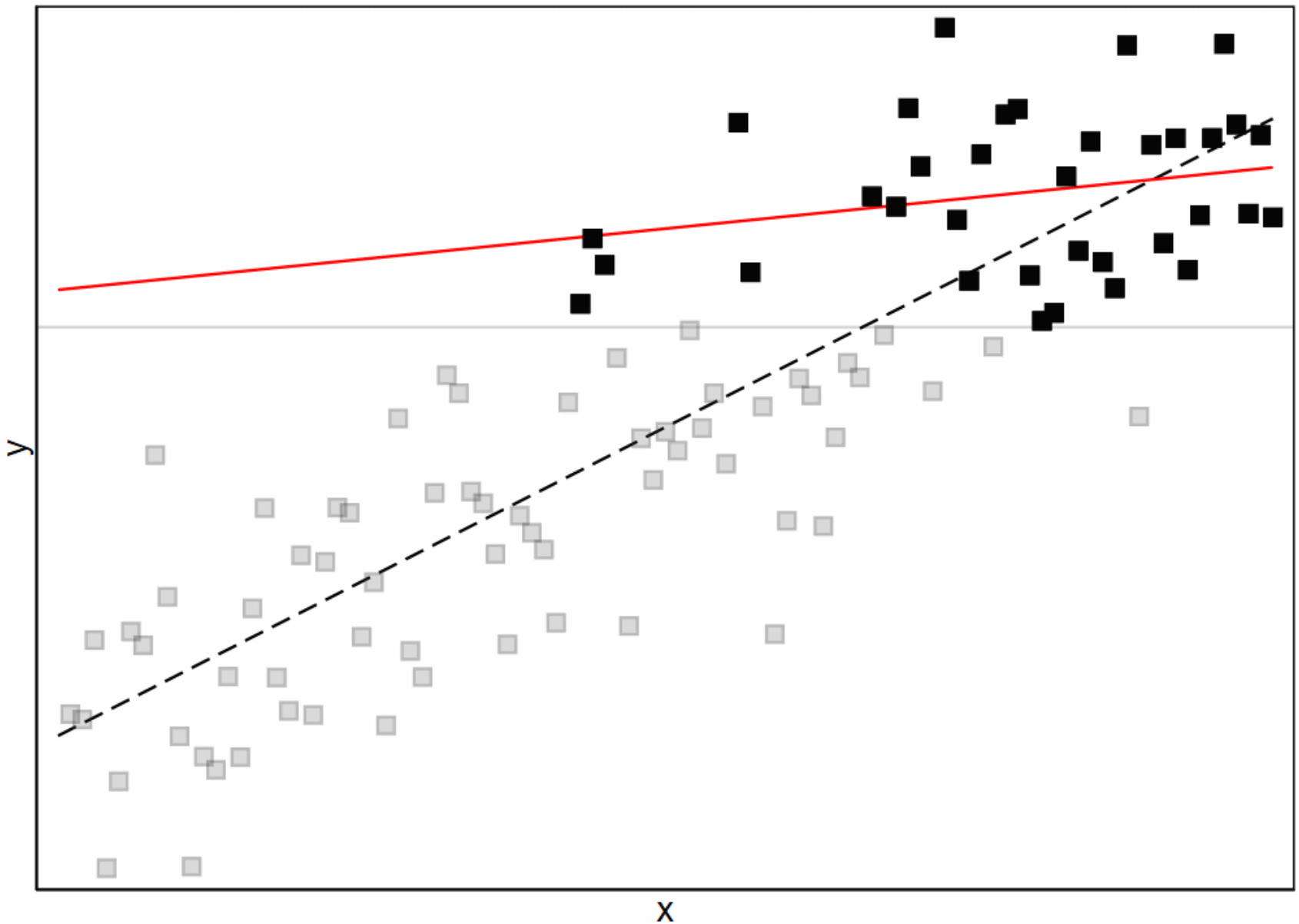
Sample selection: case 3



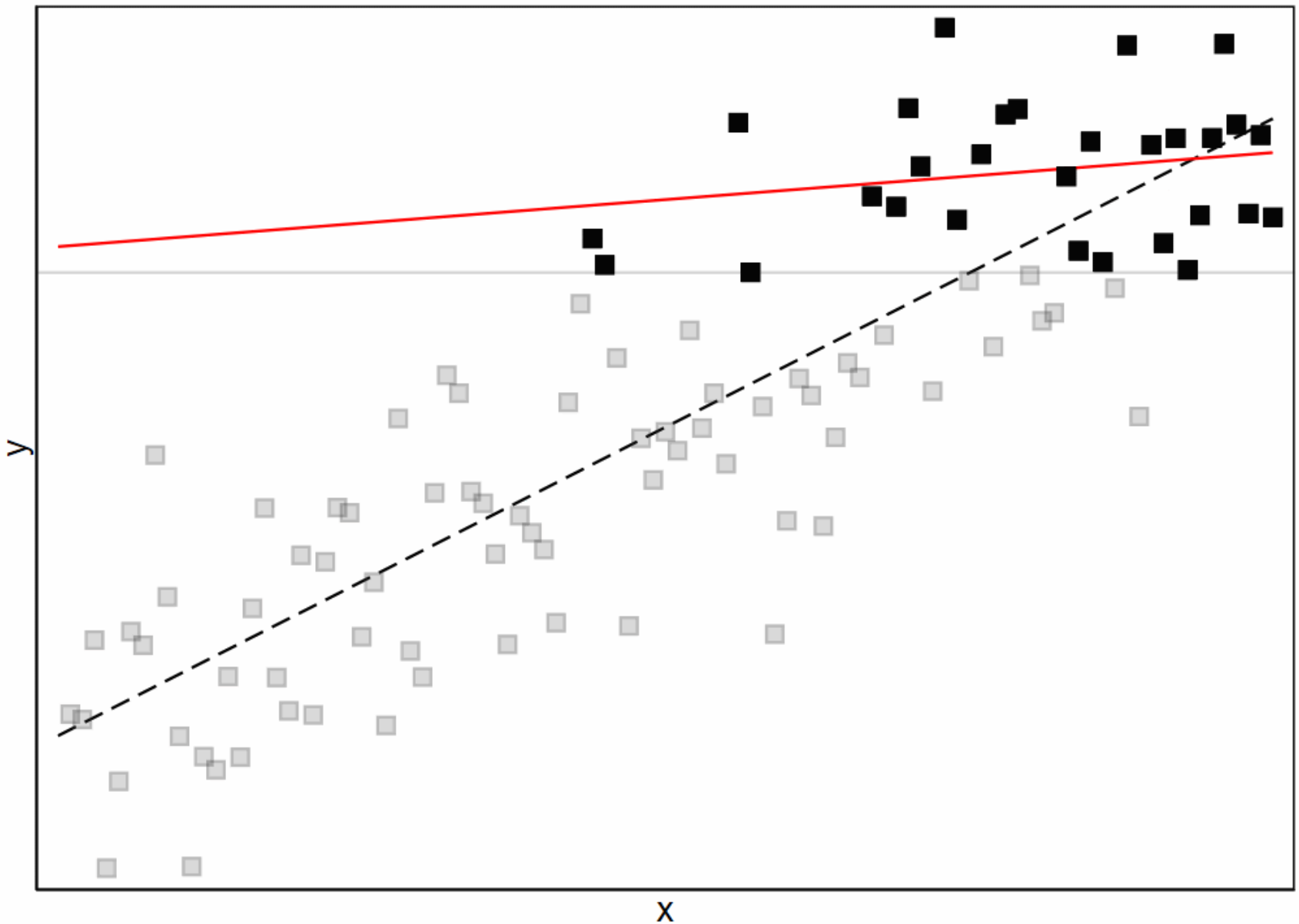
Sample selection: case 3



Sample selection: case 3



Sample selection: case 3



Types of sample selection

1. **Selection determined by the explanatory variables**
    1. Missing completely at random
    2. Missing at random
    3. Missing not at random
2. **Selection on the dependent variable**
    1. Truncated regression
    2. Incidental truncation: Heckman selection model

Linear model with missing data

A population is represented by the random vector $(\mathrm{x}, y, \mathrm{z})$

- Consider the identified IV case (generalize OLS):

$$\begin{aligned} y &= \mathrm{x}\beta + u \\ E(\mathrm{z}'u) &= 0 \end{aligned}$$

- **Random sample**
  - *rank* $E(\mathrm{z}'\mathrm{x}) = K \Rightarrow$ 2SLS to consistently estimate β
- **Selected sample**
  - Conditions are not usually enough to consistently estimate β

Linear model with missing data

- Each observation i is supplemented by a **selection indicator** s_i
  - $s_i = 1 \Rightarrow$ observation i is USED in the estimation
  - $s_i = 0 \Rightarrow$ observation i is NOT USED in the estimation
- Our sample consists of $\{(\mathrm{x}_i, y_i, \mathrm{z}_i, s_i) : \ i = 1, ..., N\}$
- The IV estimator using the selected sample can be written as

$$\hat{\beta}_{IV} \quad = \quad \left(N^{-1} \sum_{i=1}^{N} s_i \mathrm{z}_i' \mathrm{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i \mathrm{z}_i' y_i \right)$$

- $\hat{\beta}_{IV}$ is called a **complete case estimator**

Linear model with missing data

- With a large sample:

$$plim_{N\to\infty}(\hat{\beta}_{IV}) = \beta + [E(sz'x)]^{-1}E(sz'u)$$

- Assumptions for **consistency**:
  - **Rank condition**

$$rank\ E(z'x|s=1) = K$$

  - **Exogeneity or orthogonality condition**

$$E(sz'u) = P(s=1)E(z'x|s=1) = 0$$

- When do we achieve these conditions?

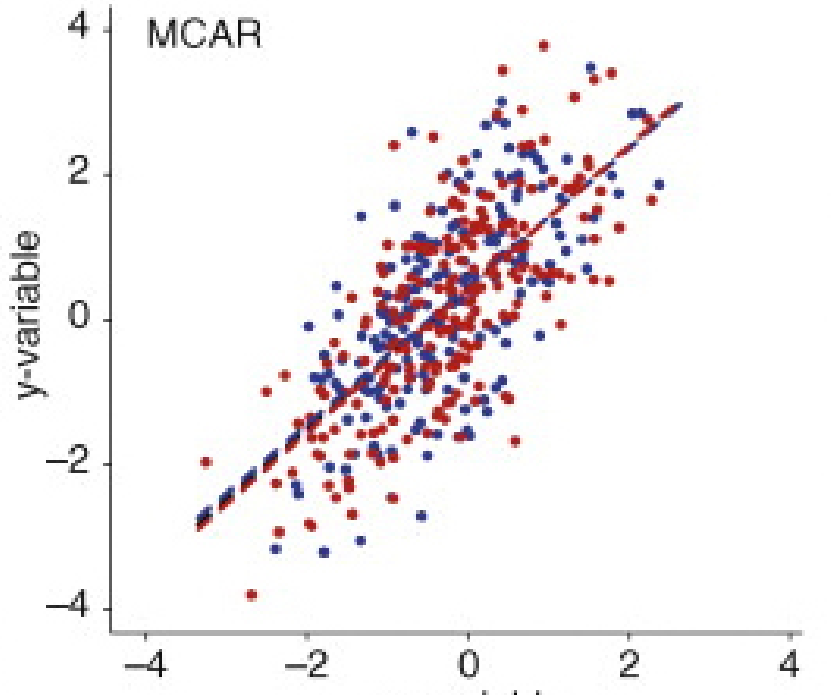
Missing completely at random (MCAR)

MCAR: s is independent of (x, y, z)

- Under MCAR

$$E(sz'u) = E(s)E(z'u) = \rho \cdot 0 = 0$$

- Selection does not affect identifying assumptions $\Rightarrow$ **OLS using the $s_i = 1$ observations is consistent for β**



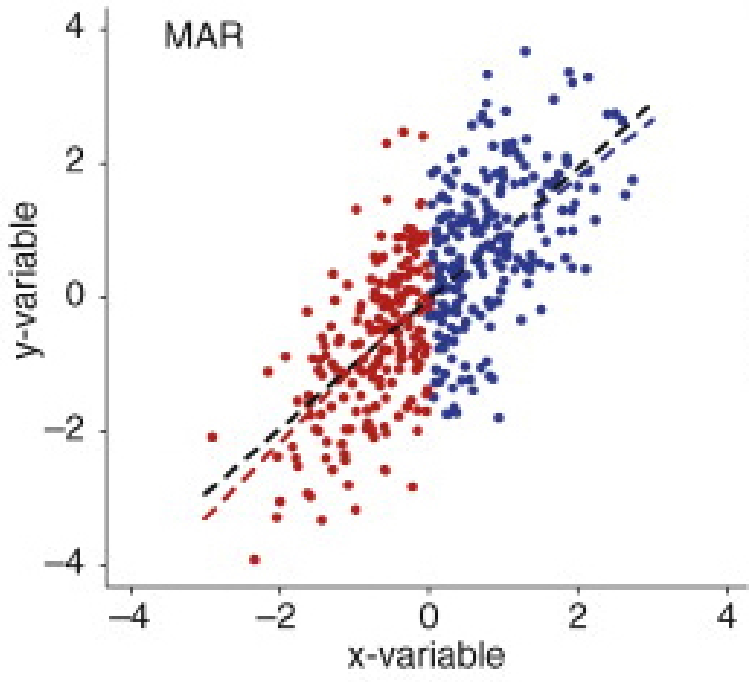
Missing at random (MAR)

MAR: $E(y|\mathrm{x}, s) = \mathrm{x}\beta$

- With MAR a sufficient condition is

$$E(y|\mathrm{x}, s) = E(y|\mathrm{x}) = \mathrm{x}\beta$$

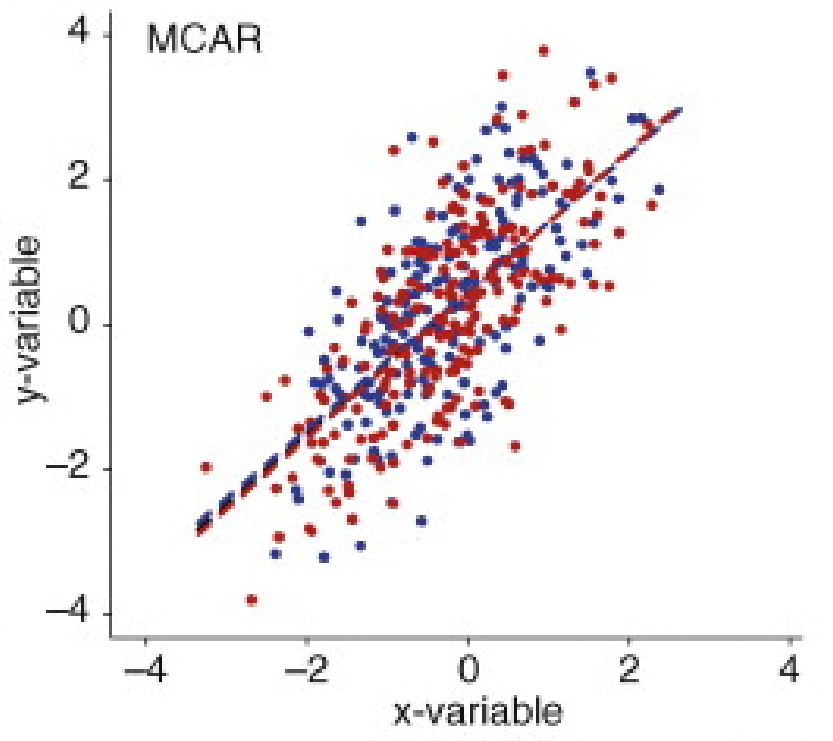
- s can be an arbitrary function of the exogenous variables.



Missing not at random (MNAR)

MNAR: previous conditions are not valid

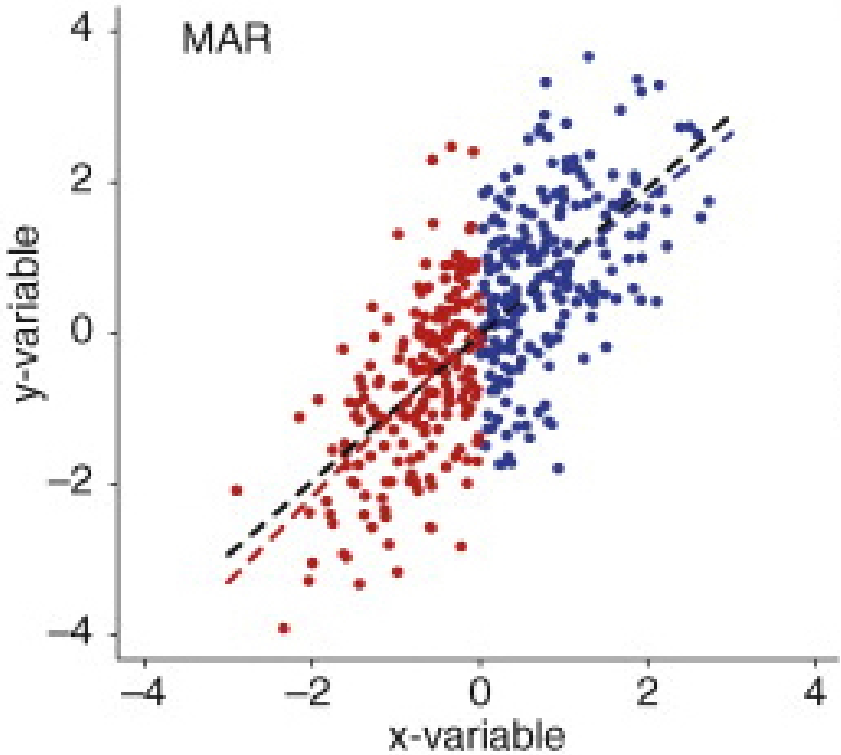
- Using MCAR and MAR assumption allow us using the complete case estimator
- Using OLS for the MNAR would lead to biased estimates



Missing not at random (MNAR)

MNAR: previous conditions are not valid

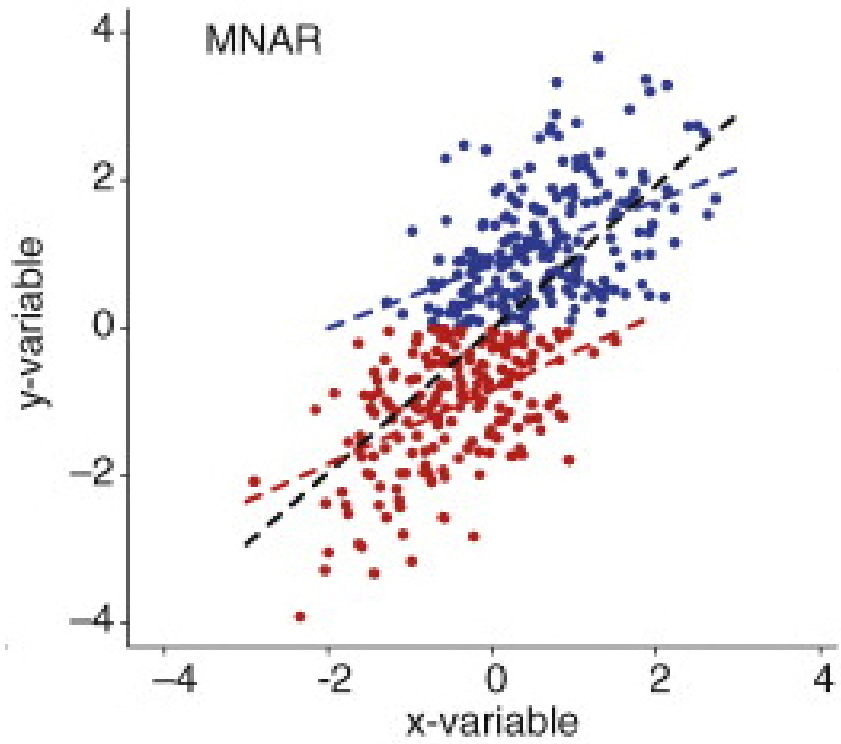
- Using MCAR and MAR assumption allow us using the complete case estimator
- Using OLS for the MNAR would lead to biased estimates



Missing not at random (MNAR)

MNAR: previous conditions are not valid

- Using MCAR and MAR assumption allow us using the complete case estimator
- Using OLS for the MNAR would lead to biased estimates



Dealing with MNAR: s correlated with u?

Suppose the population model is

$$\begin{aligned} y &= \mathrm{x}\beta + u \\ E(u|\mathrm{x}) &= 0 \\ Corr(s,u) &\neq 0 \end{aligned}$$

- Suppose
  1. s is a deterministic function of (x, v) for some variable v
  2. (u, v) is independent of x
- The conditional mean of y is

$$E(y|\mathrm{x},v) = \mathrm{x}\beta + E(u|\mathrm{x},v) = \mathrm{x}\beta + E(u|v)$$

Dealing with MNAR: s correlated with u?

We need to make additional assumptions to identify β!

- Suppose v has zero mean and $E(u|v) = \gamma v$
- The conditional mean of y can be written as

$$E(y|\mathrm{x}, v) = \mathrm{x}\beta + \gamma v$$

- Use OLS of y_i on x_i, v_i using the selected sample ($s_i = 1$) to consistently estimate β and γ
  - Notice that all variables only need to be observed when $s_i = 1$
  - **Intuition**: selection is like an omitted variable

Selection on y: truncated regression

The rule for observing a data point depends in a known deterministic way on the dependent variable

- Setting:
  1. Random draw (x_i, y_i)
  2. We only observe the data point if $s_i = 1$ for known constants a_1 and a_2

$$s_i = 1[a_1 < y_i < a_2]$$

- Allow for the cases $a_1 = -\infty$ and $a_2 = +\infty$ to have truncation only in one side.
- OLS on the selected sample $\Rightarrow \beta$ is inconsistent because selection is a function of y_i

Density conditional on $s = 1$

Assume that the population conditional density is $f(y|\mathrm{x};\beta)$ where γ is another set of parameters.

- The density conditional on $s = 1$ is

$$p(y|\mathrm{x}, s = 1) = \frac{f(y|\mathrm{x};\beta)}{P(a_1 < y < a_2|\mathrm{x})} = \frac{f(y|\mathrm{x};\beta)}{F(a_2|\mathrm{x};\beta) - F(a_1|\mathrm{x};\beta)}$$

- MLE $\rightarrow$ log-likelihood function for the selected subpopulation

$$\sum_{i=1}^{N} \{\log[f(y_i|\mathrm{x}_i;\beta)] - \log[F(a_2|\mathrm{x}_i;\beta) - F(a_1|\mathrm{x}_i;\beta)]\}$$

Truncated normal regression model

Truncated normal regression (or Truncated Tobit) model makes the following distributional assumption:

$$D(y|\mathrm{x}) = \mathcal{N}(\mathrm{x}\beta, \sigma^2)$$

- As with censoring, truncating the sample is costly
  - We are interested in $E(y|\mathrm{x}) = \mathrm{x}\beta$ in the entire population
  - We need to specify all of $D(y|\mathrm{x})$!
- Differs from the censored normal regression model
  - No information on units not in the subpopulation with $a_1 < y < a_2$

Example: $a_1 = -\infty$

Compare the likelihood in the **truncation versus censoring** cases

- **Truncated case**

$$\left\{ \frac{\sigma^{-1}\phi[(y_i - \mathrm{x}_i\beta)/\sigma]}{\Phi[(a_2 - \mathrm{x}_i\beta)/\sigma]} \right\}^{s_i}$$

  - Completely drop all units with $s_i = 0$

- **Censored case**

$$\{\sigma^{-1}\phi[(y_i - \mathrm{x}_i\beta)/\sigma]\}^{s_i}\{1 - \Phi[(a_2 - \mathrm{x}_i\beta)/\sigma]\}^{1-s_i}$$

  - Uses additional information from the binary selection indicator

- **If you have a choice, you should use censored regression**

Incidental truncation: self-selection

Sample selection is **not a deterministic function** of x or y, but it may be related to them.

- **Self-selection**: y is observed only when a certain event is true.
  - The event is often a choice of the individual we observe.
- **Example**:
  - $y = \log(w^o)$, where w^o is the "wage offer"
  - A person works if the wage offer is larger than reservation wage (w_i^r)

$$w_i^o > w_i^r$$

  - We observe $wage^o$ only if the person decides to enter the work force.
  - We do not observe $wage^o$ otherwise.

Wage offers and incidental truncation

Assume we can model the wage offer and reservation wages as

$$\begin{aligned} w_i^o &= \exp(\mathrm{x}_{i1}\beta_1 + u_{i1}) \\ w_i^r &= \exp(\mathrm{x}_{i2}\beta_2 + \gamma_2 a_i + u_{i2}) \end{aligned}$$

- Observe w_i^o if $\log(w_i^o) - \log(w_i^r) > 0$ or

$$\mathrm{x}_{i1}\beta_1 + u_{i1} - \mathrm{x}_{i2}\beta_2 - \gamma_2 a_i - u_{i2} > 0$$

  or

$$\mathrm{x}_i\delta_2 + v_{i2} > 0$$

  where x_i includes all nonredundant elements of x_{i1} and x_{i2} and a_i

- Simplest example of a **structural (econometrics) model**

General model: Type II Tobit Model

A general population model:

$$\begin{aligned} y_1 &= \mathrm{x}_1\beta_1 + u_1 \\ y_2 &= 1[\mathrm{x}\delta_2 + v_2 > 0] \end{aligned}$$

- y_1 is the **response variable** (only partially observed)
- y_2 is the **selection indicator** (what we called s before)
- **Assumptions**
  1. (x, y_2) are always observed, y_1 is observed only when $y_2 = 1$
  2. (u_1, v_2) is independent of x with u_1 having mean zero
  3. $v_2 \sim \mathcal{N}(0, 1)$
  4. $E(u_1|v_2) = \gamma_1 v_2$

An estimating equation for β_1

Under the previous assumptions

$$\begin{aligned} E(y_1|\mathrm{x}, v_2) &= \mathrm{x}_1\beta_1 + E(u_1|\mathrm{x}, v_2) \\ &= \mathrm{x}_1\beta_1 + E(u_1|v_2) \\ &= \mathrm{x}_1\beta_1 + \gamma_1 v_2 \end{aligned}$$

- **Problem**: we observe y_2, not v_2!

$$\begin{aligned} E(y_1|\mathrm{x}, y_2) &= E[E(y_1|\mathrm{x}, v_2)|\mathrm{x}, y_2] \\ &= \mathrm{x}_1\beta_1 + \gamma_1 E(v_2|\mathrm{x}, y_2) \end{aligned}$$

  - How can we recover $E(v_2|\mathrm{x}, y_2)$?

Generalized residual: $E(v_2|\mathrm{x}, y_2)$

- Decompose the term using the two values of y_2

$$E(v_2|\mathrm{x}, y_2) = y_2 E(v_2|\mathrm{x}, y_2 = 1) + (1 - y_2) E(v_2|\mathrm{x}, y_2 = 0)$$

- For the selected sample, notice that

$$\begin{aligned} E(v_2|\mathrm{x}, y_2 = 1) &= E(v_2|\mathrm{x}, v_2 > -\mathrm{x}\delta_2) \\ &= \frac{\phi(-x\delta_2)}{1 - \Phi(-x\delta_2)} \end{aligned}$$

- Reminder: step 2 comes from this property of Normal distribution (same rule as tobit):

$$E(z|a < z < b) = \mu + \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\sigma$$

Heckit method (Heckman 1976)

If we restrict to the selected sample ($y_2 = 1$) we have

$$\begin{aligned} E(y_1|\mathrm{x}, y_2 = 1) &= \mathrm{x}_1\beta_1 + \gamma_1 E(v_2|\mathrm{x}, y_2 = 1) \\ &= \mathrm{x}_1\beta_1 + \gamma_1\lambda(\mathrm{x}\delta_2) \end{aligned}$$

- **Two-step estimation method**
  1. **Selection equation**: probit of y_{i2} on x_i using all of the data

     $$\hat{\lambda}_{i2} = \lambda(\mathrm{x}_i\hat{\delta}_2)$$

  2. **Regression equation**: OLS of y_{i1} on x_{i1}, $\hat{\lambda}_{i2}$ in the selected sample
     - $H_0 : \gamma_1 = 0$ tests for no sample selection problem
     - Notice we don't need an exclusion restriction (like in IV)

APPLICATION: wage offer for married women

```
. use mroz

. des lwage inlf nwifeinc

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
lwage           float  %9.0g                  log(wage)
inlf            byte   %9.0g                  =1 if in lab frce, 1975
nwifeinc        float  %9.0g                  (faminc - wage*hours)/1000
. sum lwage inlf educ kidslt6 nwifeinc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       lwage |       428    1.190173    .7231978  -2.054164   3.218876
        inlf |       753    .5683931    .4956295          0          1
        educ |       753    12.28685    2.280246          5         17
     kidslt6 |       753    .2377158     .523959          0          3
    nwifeinc |       753    20.12896     11.6348  -.0290575         96
```

APPLICATION: wage equation using OLS

$$log(w_i) = \beta_0 + \beta_1 \cdot educ_i + \beta_2 \cdot exper_i + \beta_3 \cdot expersq_i + u$$

```
. reg lwage educ exper expersq

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  3,   424) =   26.29
       Model |  35.0222967     3  11.6740989            Prob > F      =  0.0000
    Residual |  188.305144   424  .444115906           R-squared     =  0.1568
-------------+------------------------------           Adj R-squared =  0.1509
       Total |  223.327441   427  .523015084           Root MSE      =  .66642

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1074896   .0141465     7.60   0.000     .0796837    .1352956
       exper |   .0415665   .0131752     3.15   0.002     .0156697    .0674633
     expersq |  -.0008112   .0003932    -2.06   0.040    -.0015841   -.0000382
       _cons |  -.5220406   .1986321    -2.63   0.009    -.9124667   -.1316144
------------------------------------------------------------------------------
```

APPLICATION: Heckit regression equation

```
. heckman lwage educ exper expersq, select(inlf = educ exper expersq nwifeinc
          age kidslt6 kidsge6) twostep

Heckman selection model -- two-step estimates   Number of obs      =       753
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       428

                                                Wald chi2(6)       =    180.10
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1090655    .015523     7.03   0.000     .0786411      .13949
       exper |   .0438873   .0162611     2.70   0.007     .0120163    .0757584
     expersq |  -.0008591   .0004389    -1.96   0.050    -.0017194    1.15e-06
       _cons |  -.5781032   .3050062    -1.90   0.058    -1.175904     .019698
-------------+----------------------------------------------------------------
```

APPLICATION: Heckit selection equation

inlf						
educ	.1309047	.0252542	5.18	0.000	.0814074	.180402
exper	.1233476	.0187164	6.59	0.000	.0866641	.1600311
expersq	-.0018871	.0006	-3.15	0.002	-.003063	-.0007111
nwifeinc	-.0120237	.0048398	-2.48	0.013	-.0215096	-.0025378
age	-.0528527	.0084772	-6.23	0.000	-.0694678	-.0362376
kidslt6	-.8683285	.1185223	-7.33	0.000	-1.100628	-.636029
kidsge6	.036005	.0434768	0.83	0.408	-.049208	.1212179
_cons	.2700768	.508593	0.53	0.595	-.7267473	1.266901
mills						
lambda	.0322619	.1336246	0.24	0.809	-.2296376	.2941613
rho	0.04861					
sigma	.66362875					
lambda	.03226186	.1336246				

APPLICATION: Heckit selection equation

inlf						
educ	.1309047	.0252542	5.18	0.000	.0814074	.180402
exper	.1233476	.0187164	6.59	0.000	.0866641	.1600311
expersq	-.0018871	.0006	-3.15	0.002	-.003063	-.0007111
nwifeinc	-.0120237	.0048398	-2.48	0.013	-.0215096	-.0025378
age	-.0528527	.0084772	-6.23	0.000	-.0694678	-.0362376
kidslt6	-.8683285	.1185223	-7.33	0.000	-1.100628	-.636029
kidsge6	.036005	.0434768	0.83	0.408	-.049208	.1212179
_cons	.2700768	.508593	0.53	0.595	-.7267473	1.266901
mills						
lambda	.0322619	.1336246	0.24	0.809	-.2296376	.2941613
rho	0.04861					
sigma	.66362875					
lambda	.03226186	.1336246				

Coefficient on the selection term in the second stage

APPLICATION: Heckit selection equation

inlf						
educ	.1309047	.0252542	5.18	0.000	.0814074	.180402
exper	.1233476	.0187164	6.59	0.000	.0866641	.1600311
expersq	-.0018871	.0006	-3.15	0.002	-.003063	-.0007111
nwifeinc	-.0120237	.0048398	-2.48	0.013	-.0215096	-.0025378
age	-.0528527	.0084772	-6.23	0.000	-.0694678	-.0362376
kidslt6	-.8683285	.1185223	-7.33	0.000	-1.100628	-.636029
kidsge6	.036005	.0434768	0.83	0.408	-.049208	.1212179
_cons	.2700768	.508593	0.53	0.595	-.7267473	1.266901
mills						
lambda	.0322619	.1336246	0.24	0.809	-.2296376	.2941613
rho	0.04861					
sigma	.66362875					
lambda	.03226186	.1336246				

Coefficient on the selection term in the second stage is not significant