Quantitative analysis +

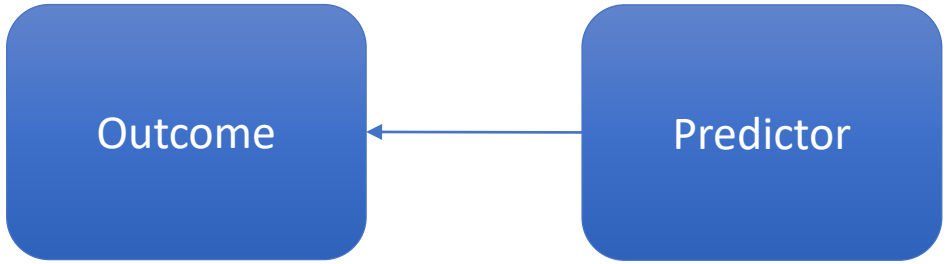
Chris Moreh, 2024

Week 2 Escaping Flatland

Linear models and their limitations

Models

- Descriptive
- Predictive
- Inferential
- Causal



Models

- Descriptive
- Predictive
- Inferential
- Causal

- The aim of modelling is to explain
- Are we assuming a “linear” relationship?
- Are we assuming that no other factors affect trust?
- Are we assuming that the relationship between “inequality” and trust is not affected by other factors? (hidden “third” variables?)

Inference

- From “data” to “population”
- From “association” to “causality”
- ***Statistical inference*** can be formulated as a set of operations on data that yield ***estimates*** and ***uncertainty statements*** about ***predictions*** and ***parameters*** of some underlying ***process*** or ***population*** (Gelman, Hill, and Vehtari 2020)
- From a mathematical standpoint, these *probabilistic uncertainty statements* are derived based on some ***assumed probability model*** for observed data.
- The ***normal (Gaussian) distribution*** — *linear regression*
- The ***binomial distribution*** — *logistic regression*
- So far we have focused on the “point” **estimates** from regression models

Estimation

Terminology

- Outcome: y
- Predictor: x
- Observed y, y: truth
- Predicted y, $\hat{y}$: fitted, estimated
- Residual: difference between observed and predicted outcome for a given value of predictor

Model evaluation

- One concern in evaluating models is how well they do for prediction
- We're generally interested in how well a model might do for predicting the outcome for a new observation, not for predicting the outcome for an observation we used to fit the model (and already know its observed value)

Linear regression

What is simple (bivariate) linear regression?

- **Simple**: it involves one *response* (dependent; outcome) variable and only one *explanatory* (independent; predictor) variable
- **Linear**: we assume that a one-unit change in the independent variable leads to a certain amount of increase or decrease in the dependent variable
- **Regression**: the term originated in Francis Galton's studies of biological inheritance, specifically his 1886 article “Regression towards mediocrity in hereditary stature", published in the *Journal of the Anthropological Institute of Great Britain and Ireland* (vol. 15, pp. 246–263)
- At its simplest, it helps us evaluate whether there is a linear relationship between a *numerical* variable on a horizontal axis and the average of a *numerical* variable on the vertical axis. It provides a mathematical solution to the question: what is the *best fitting* straight line to capture the relationship of two variables in a *scatter plot*?

Linear regression

What is simple (bivariate) linear regression?

- A statistical method for fitting a line to data where the relationship between two variables, x and y, can be modeled by a straight line with some error:

$$\overbrace{\mathbf{y}}^{outcome} = b_0 + b_1 * \overbrace{\mathbf{x}}^{predictor} + e$$

- A dataset normally contains a collection of values for y and x from a number (N) of cases (e.g. survey respondents, or another unit of analysis)
- For example, each row i in a dataset containing $N = 1 \ldots n$ cases/units/rows would store values for y_i and x_i
- so that:

$$y_i = b_0 + b_1 x_i + e_i$$

Linear regression

What is simple (bivariate) linear regression?

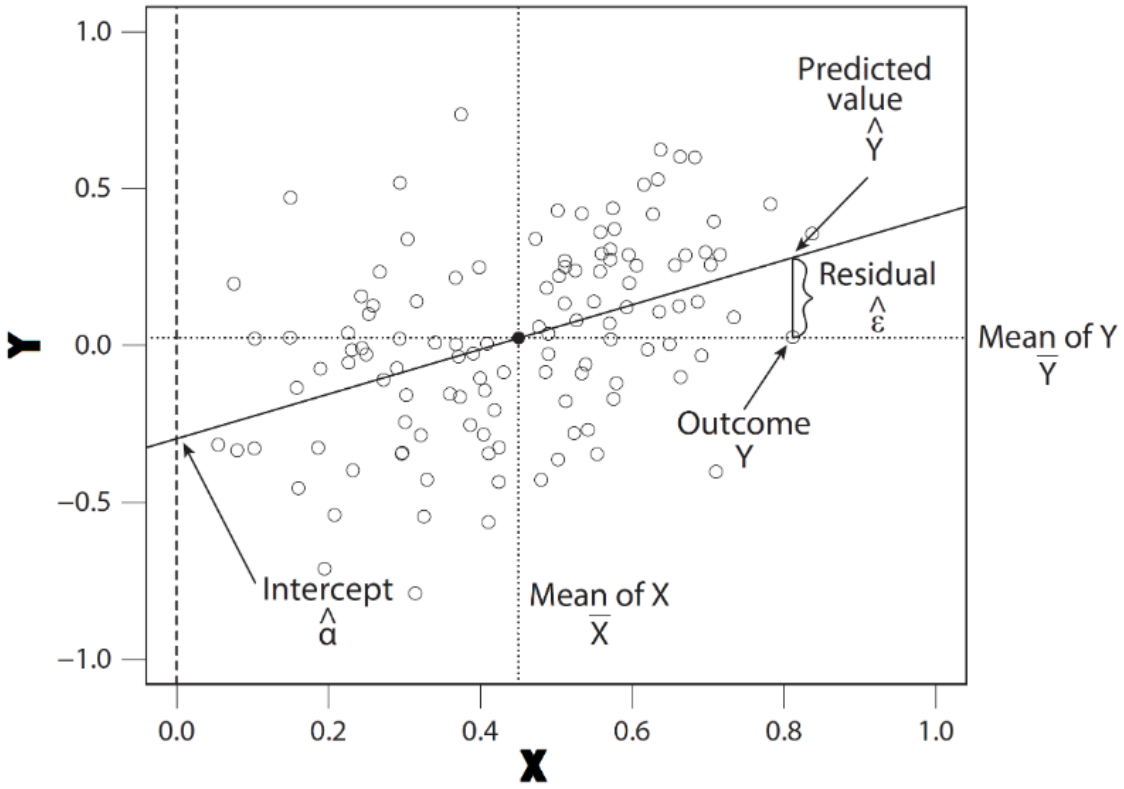
$$\overbrace{\mathbf{y}}^{outcome} = \overbrace{b_0}^{intercept} + \overbrace{b_1}^{slope} * \overbrace{\mathbf{x}}^{predictor} + \overbrace{e}^{error;\ residual}$$

- our aim is to calculate values for the b_0 and b_1 based on our data
- the intercept b_0 (or α, as it is also sometimes notated) represents the average value of y when x is zero
- the slope b_1 measures the average increase in y when x increases by one unit
- together, b_0 and b_1 together are called *coefficients*
- the error term e allows an observation to deviate from a perfect linear relationship
- More generally, our aim is to generalise from our *sample* data to a hypothetical *population*. The mathematical notation to describe this general case is:

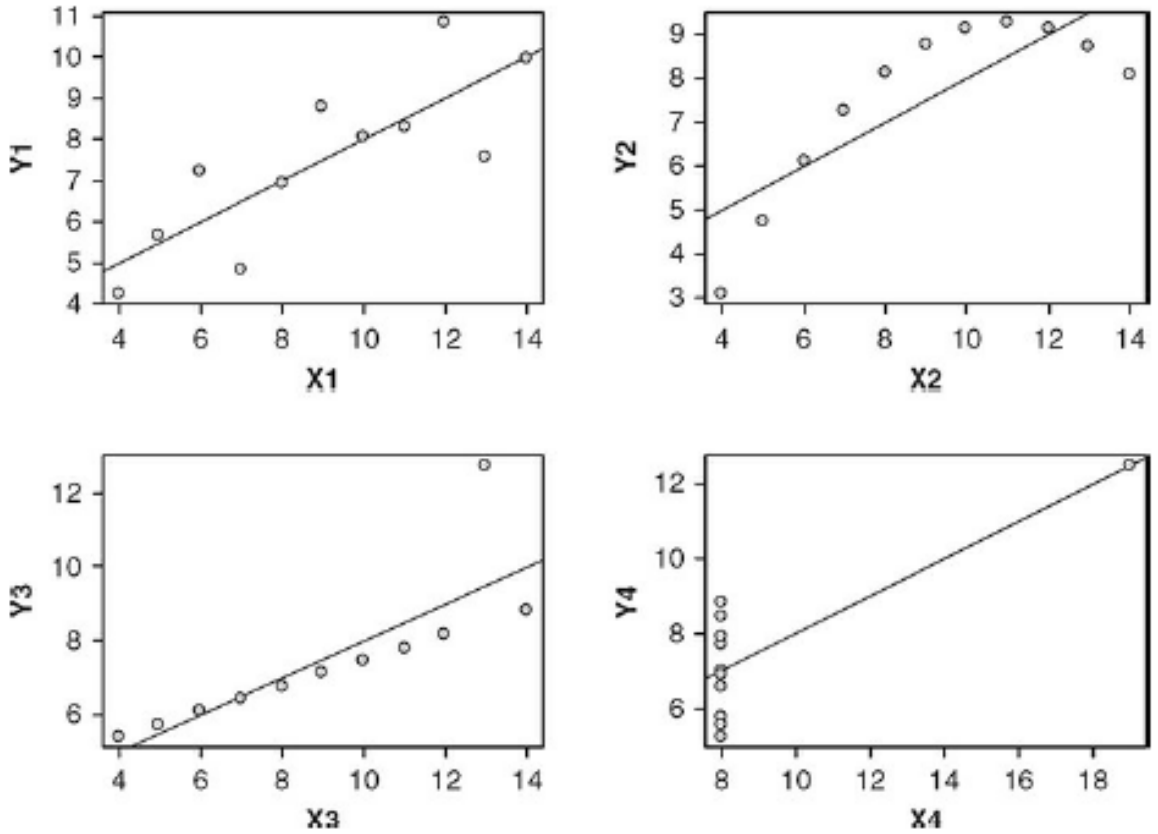
$$Y = \beta_0 + \beta_1 * X + \epsilon$$

Linear regression

For example, say we have data on two variables, X and Y; both are continuous measures with means of 0.48 and 0.006, respectively, and some standard deviation (i.e. they have variation among values):



Linear regression

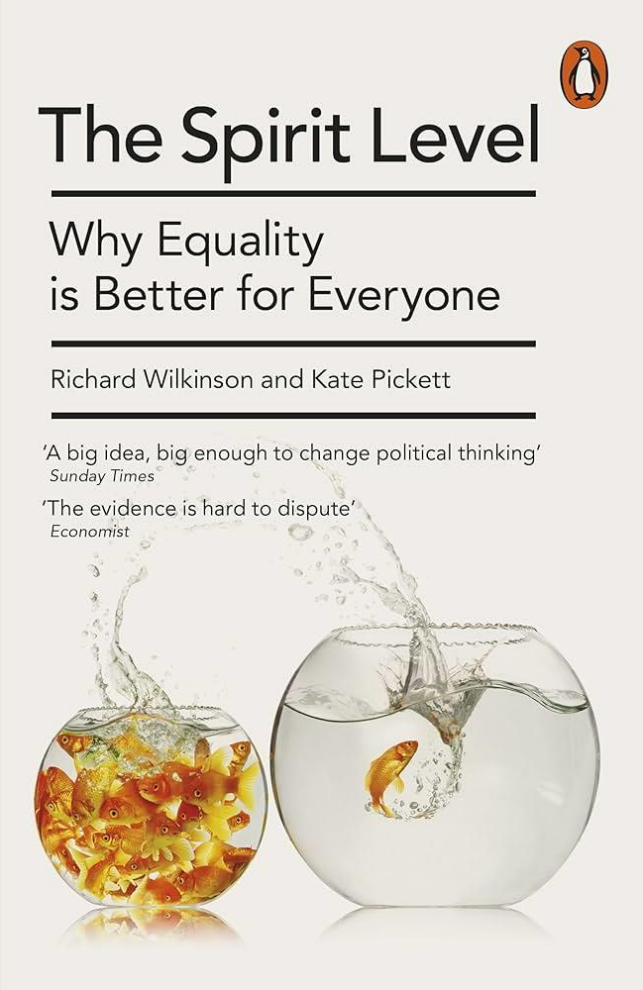


Source: F. J. Anscombe (1973). Graphs in Statistical Analysis. *The American Statistician*, 27(1):17-21. If you have access to JSTOR you can get the article at the following link: http://www.jstor.org/stable/2682899

Line of "best fit"?

The same regression line could represent very different relationships

“The spirit level” (2010)



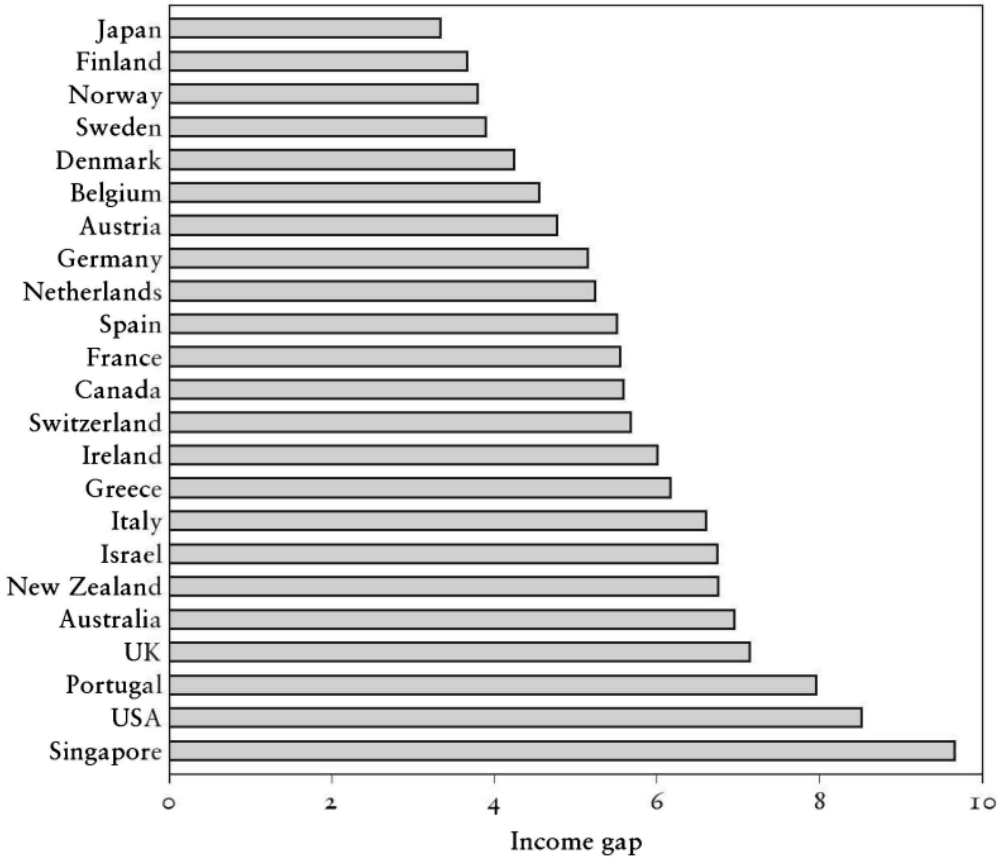
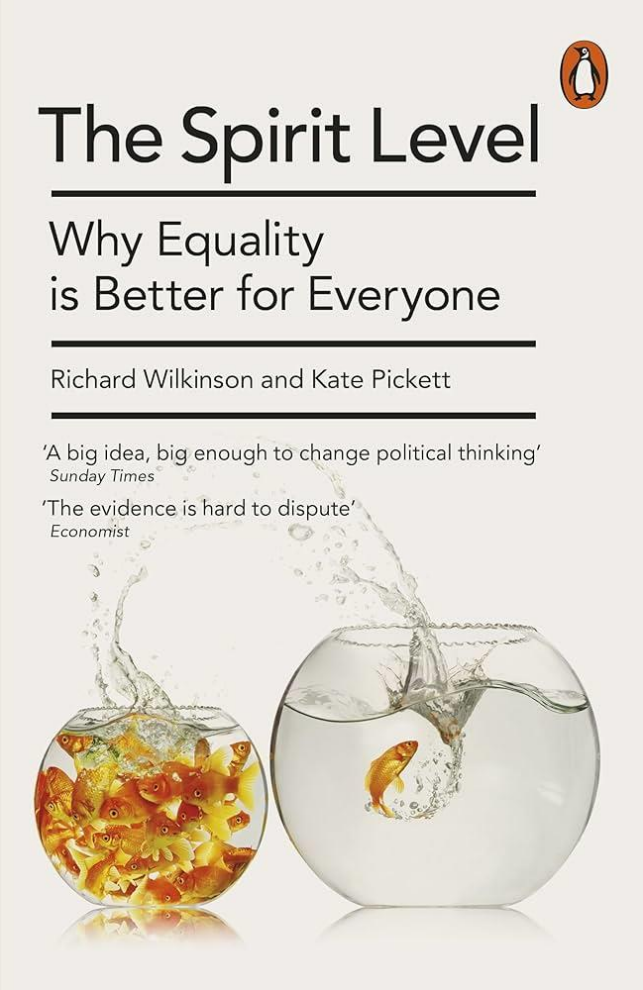


Figure 2.1 *How much richer are the richest 20 per cent than the poorest 20 per cent in each country?*[2]

“The spirit level” (2010)



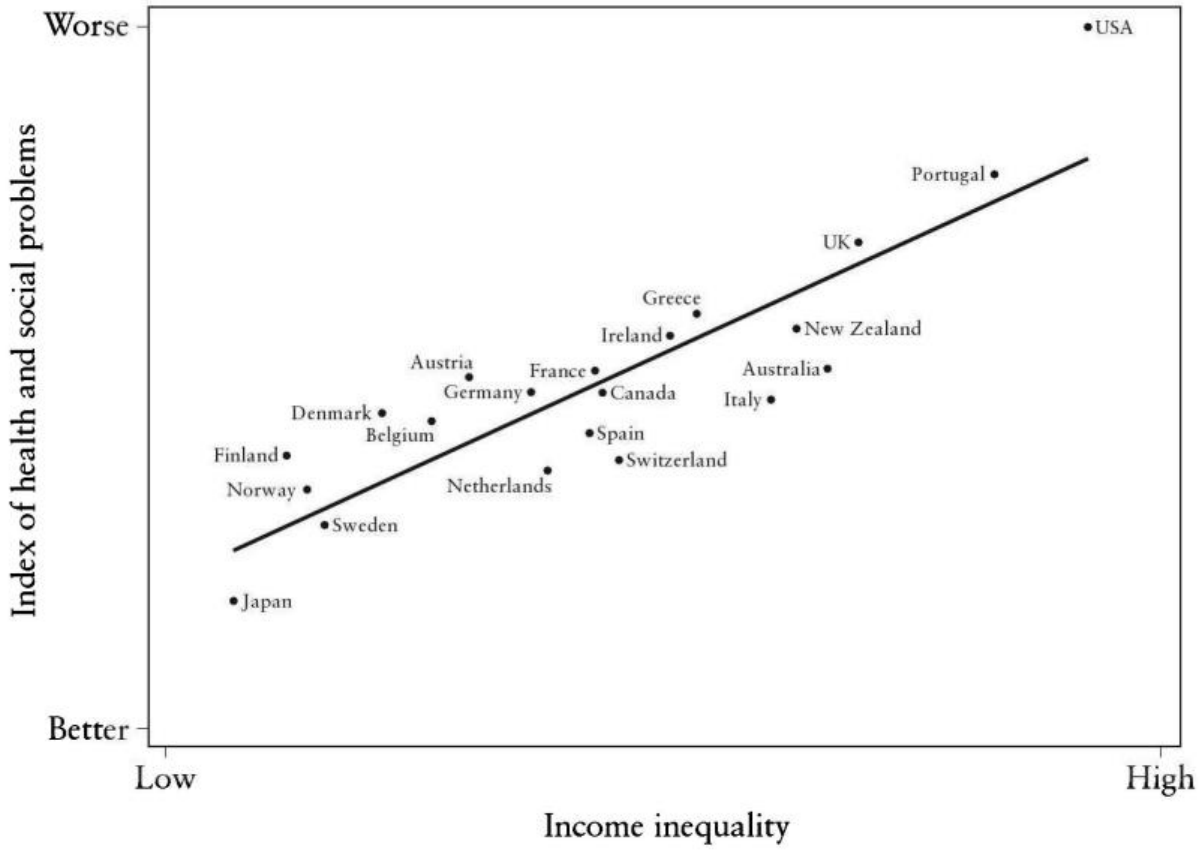


Figure 2.2 *Health and social problems are closely related to inequality among rich countries.*

“The spirit level” (2010)

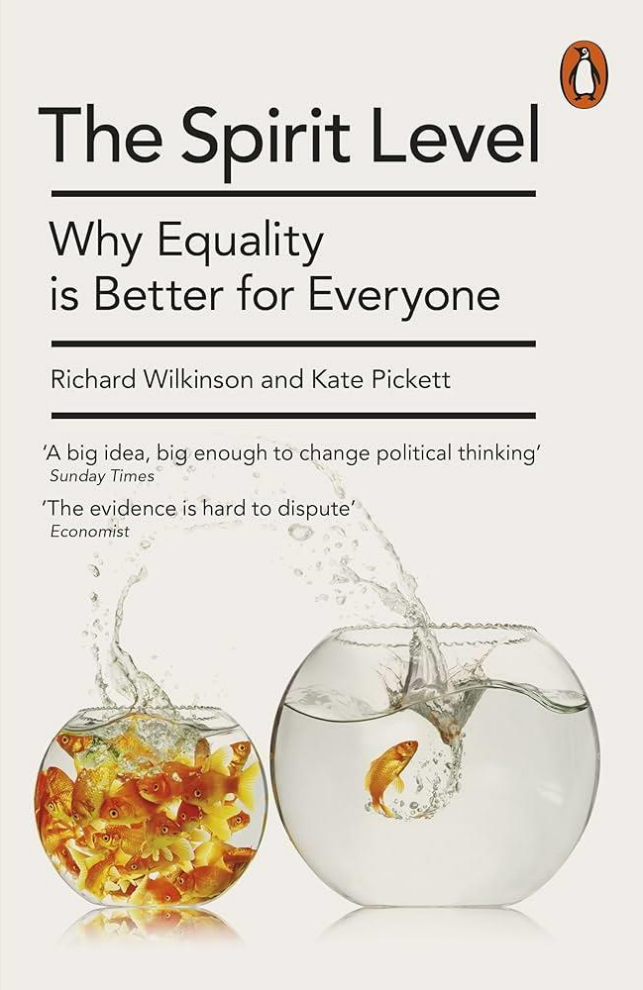


Figure 4.1 *The percentage of people agreeing that ‘most people can be trusted’ is higher in more equal countries.*

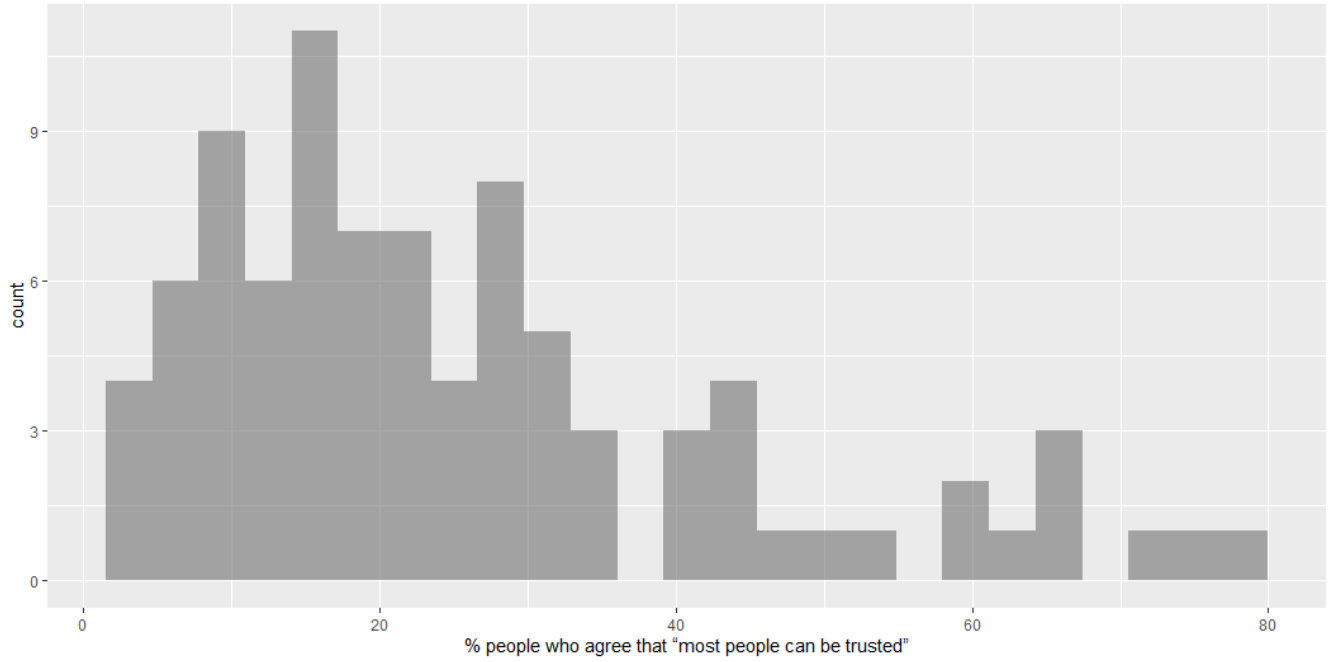
New data (WVS + World Bank)

```
describe_distribution(inequality)
```

Variable	Mean	SD	IQR	Range	Skewness	Kurtosis	n	n_Missing
trust_pct	25.73	18.41	21.14	[2.14, 77.42]	1.13	0.62	89	0
GDPpercap2	27447.01	21342.56	28284.55	[1987.97, 1.23e+05]	1.62	4.20	86	3
pop	5.78e+07	1.58e+08	4.68e+07	[73837.00, 1.40e+09]	7.32	61.13	87	2
urban_pop_pct	67.92	18.93	25.93	[20.31, 100.00]	-0.49	-0.38	87	2
inc_top20	42.69	5.09	6.80	[34.56, 57.35]	0.86	0.42	79	10
inc_bottom20	7.15	1.62	2.32	[3.45, 10.08]	-0.19	-0.69	79	10
s80s20	6.50	2.60	2.79	[3.52, 16.64]	1.67	3.50	79	10

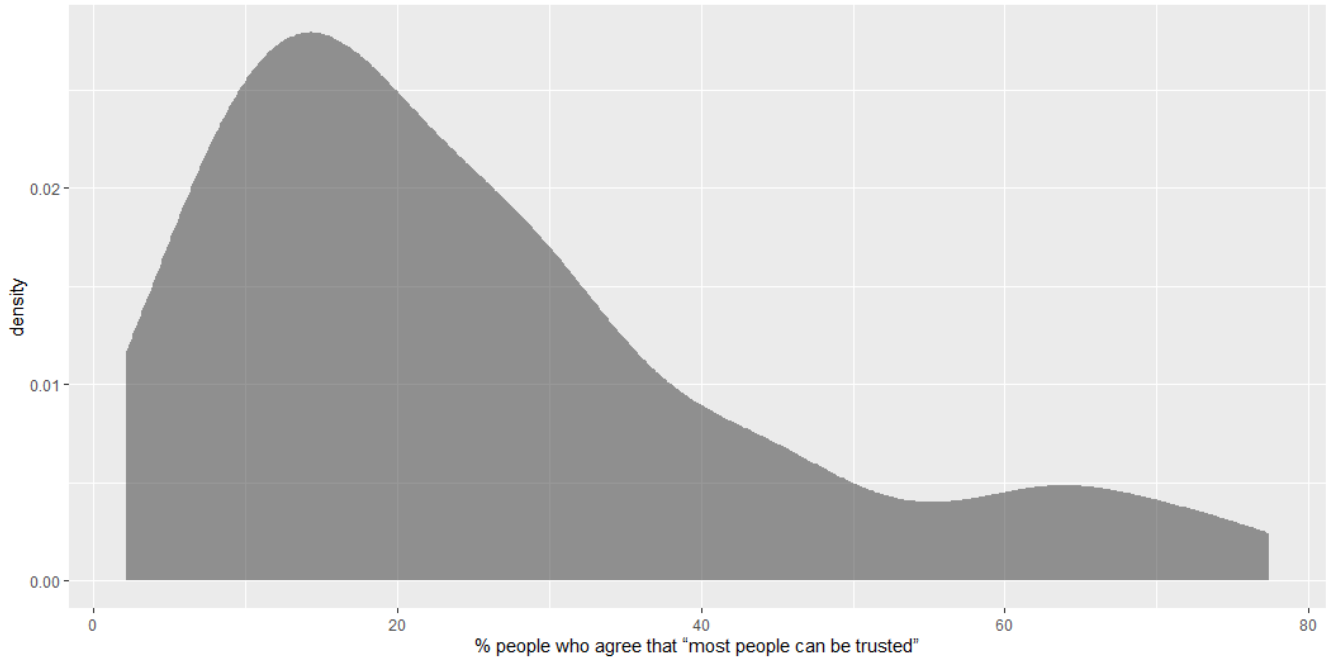
New data (WVS + World Bank)

```
gf_histogram( ~ trust_pct,
data = inequality)
```



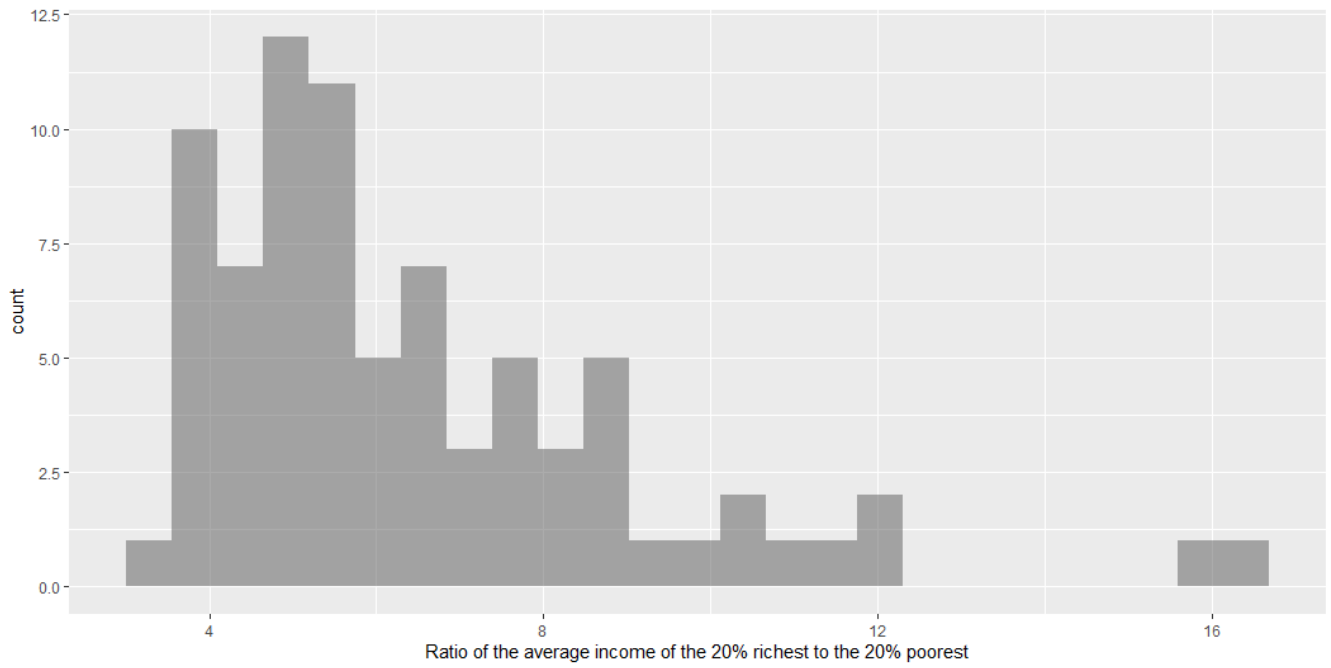
New data (WVS + World Bank)

```
gf_density( ~ trust_pct, data
= inequality)
```

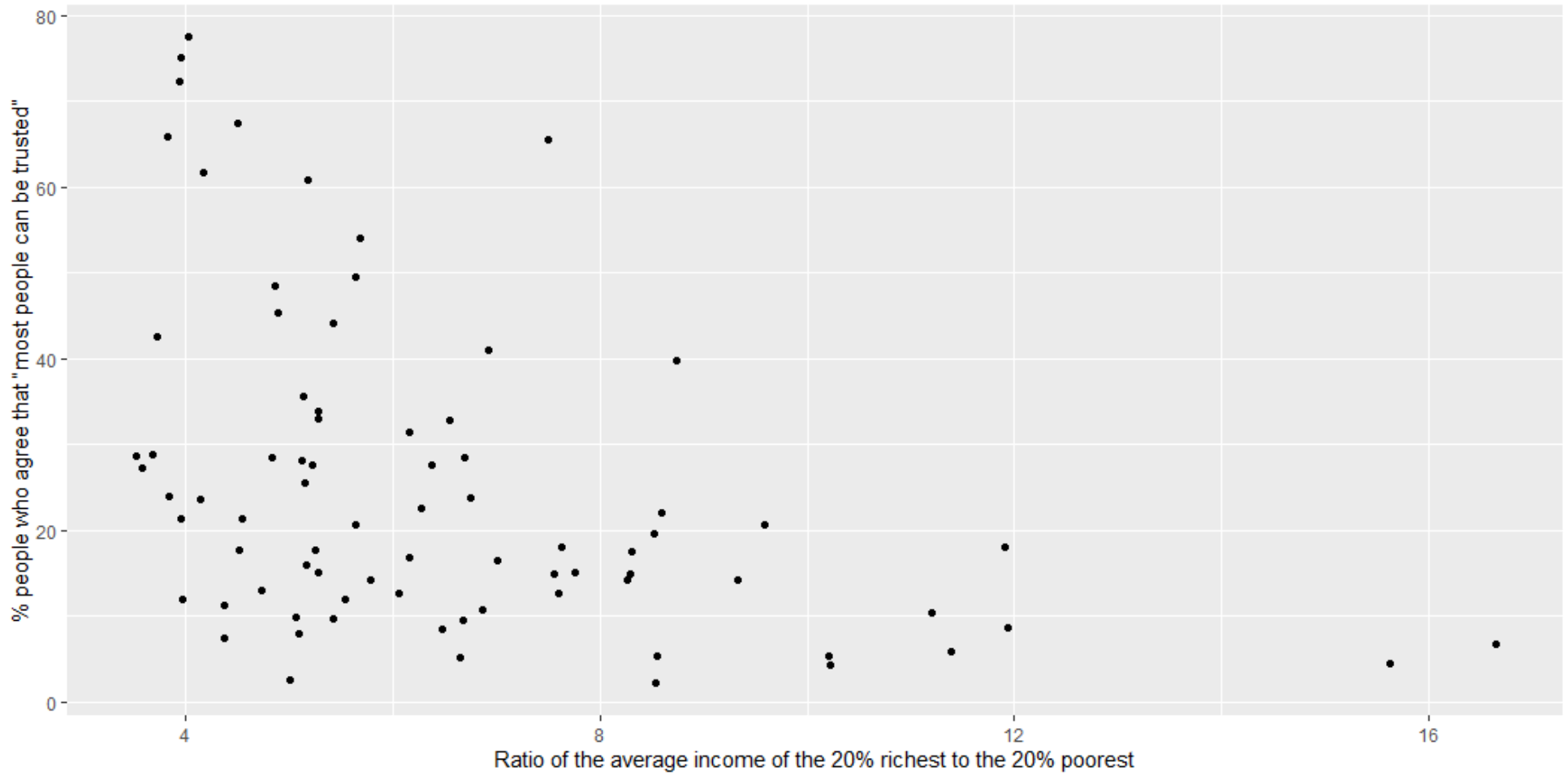


New data (WVS + World Bank)

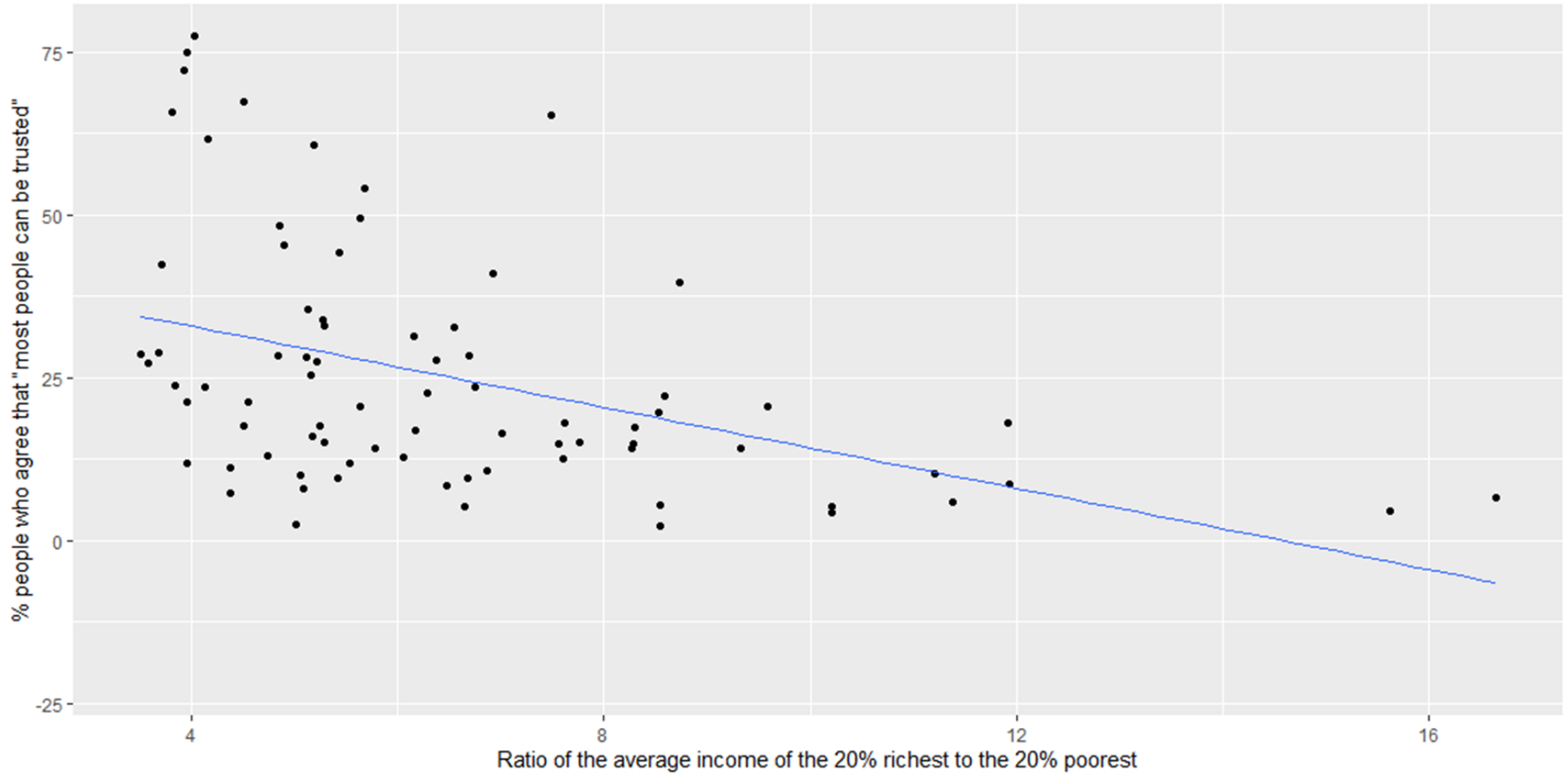
```
gf_histogram( ~ s80s20,
data = inequality)
```



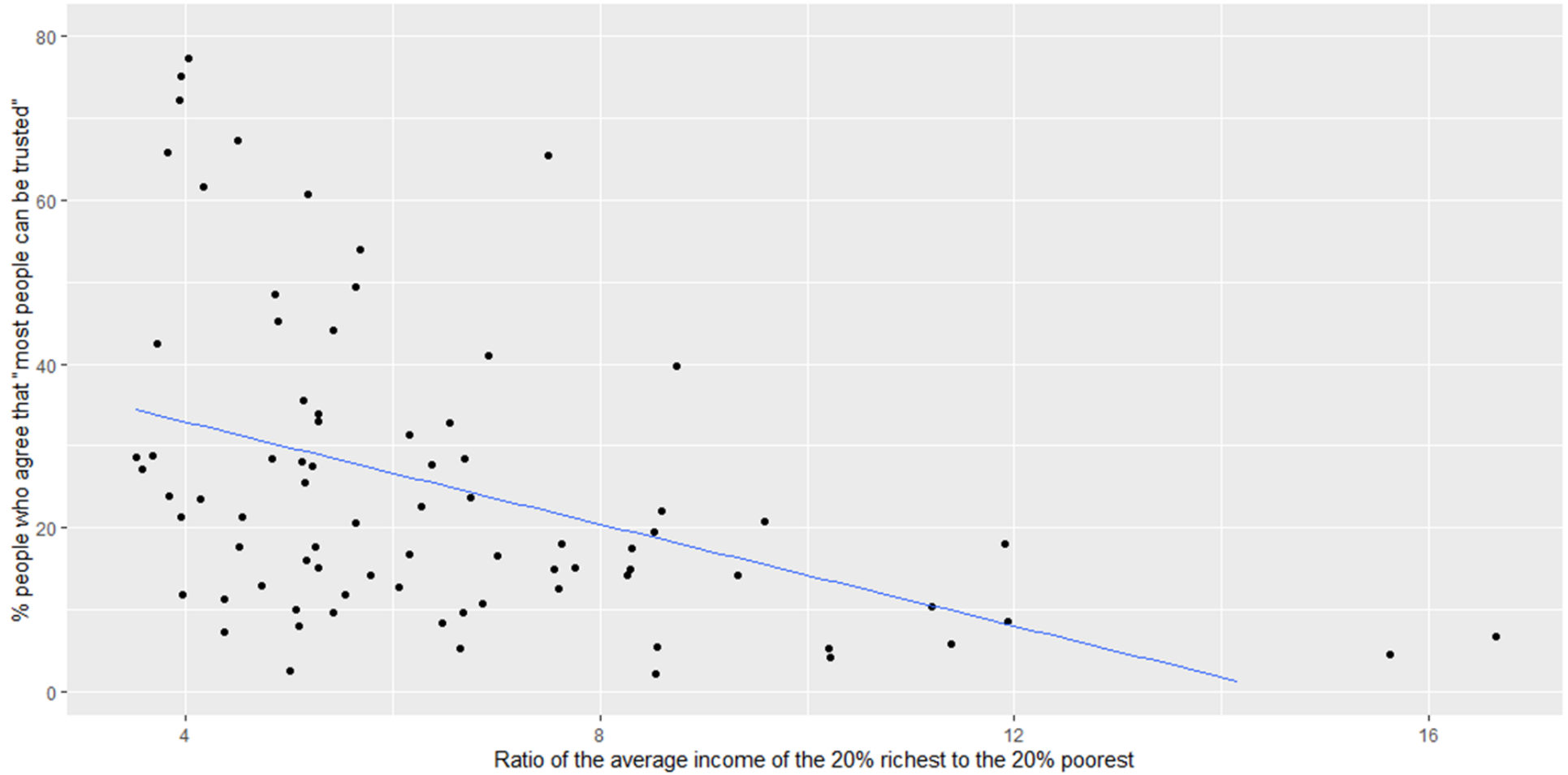
New data (WVS + World Bank)



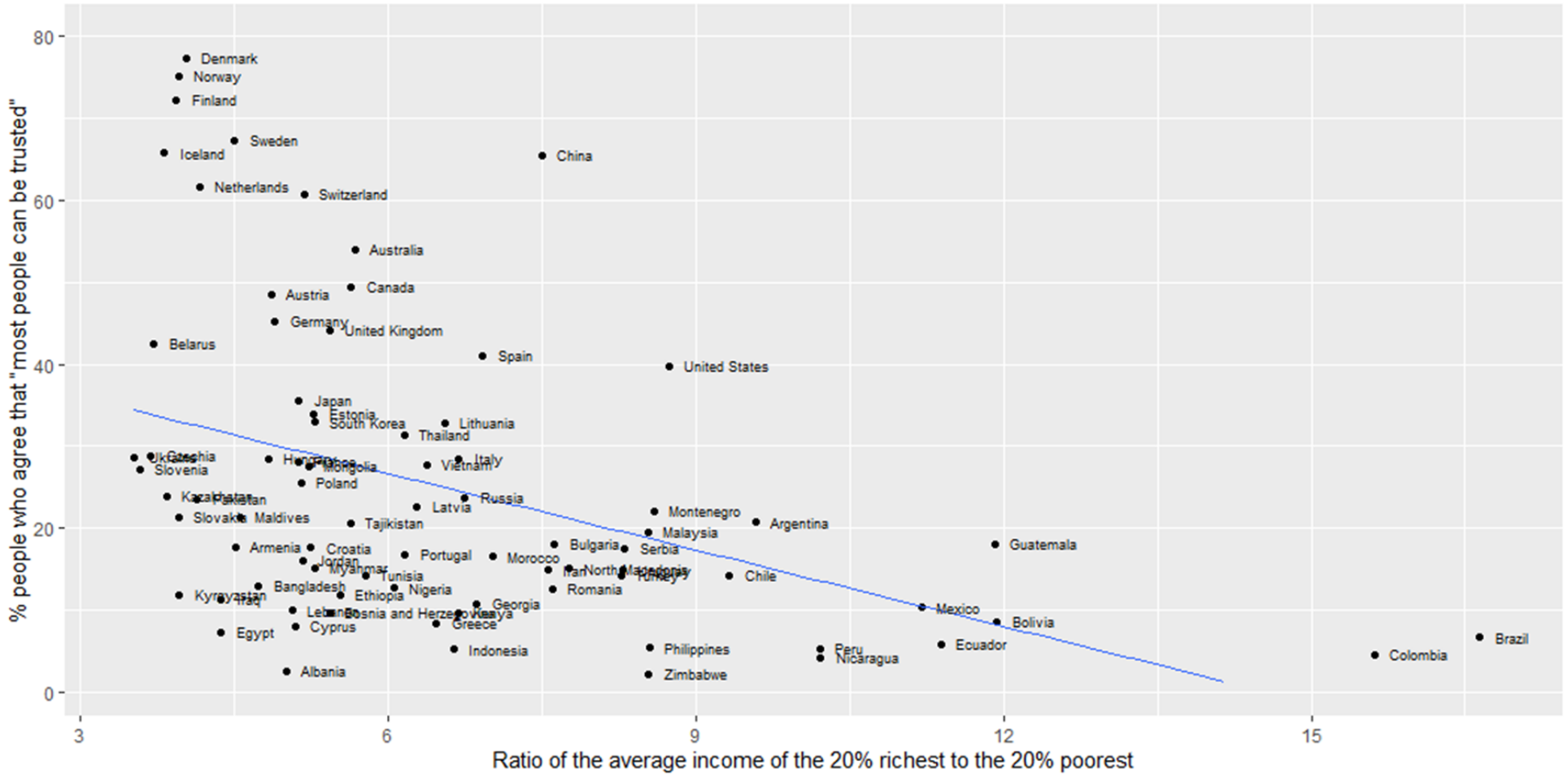
New data (WVS + World Bank)



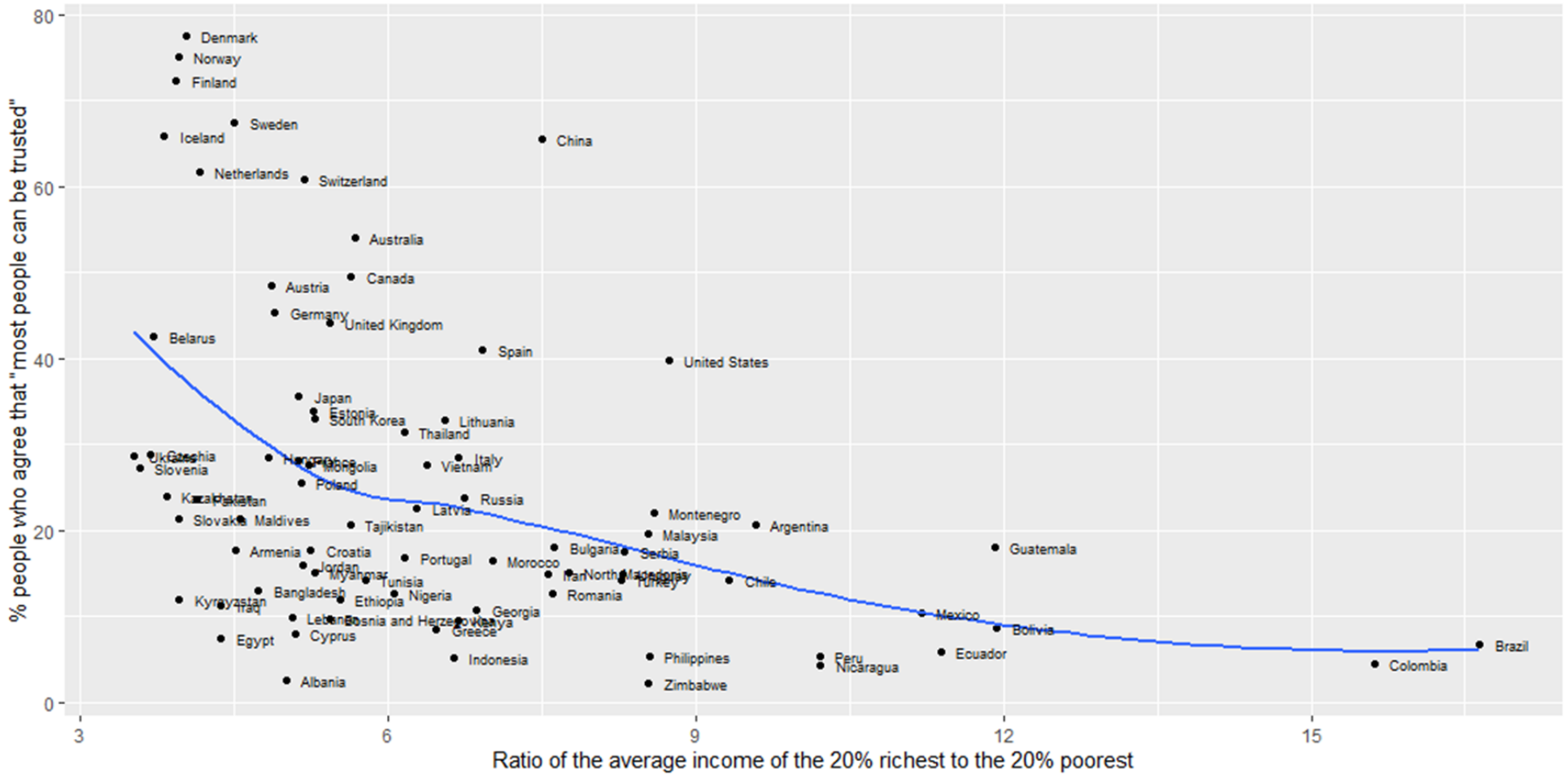
New data (WVS + World Bank)



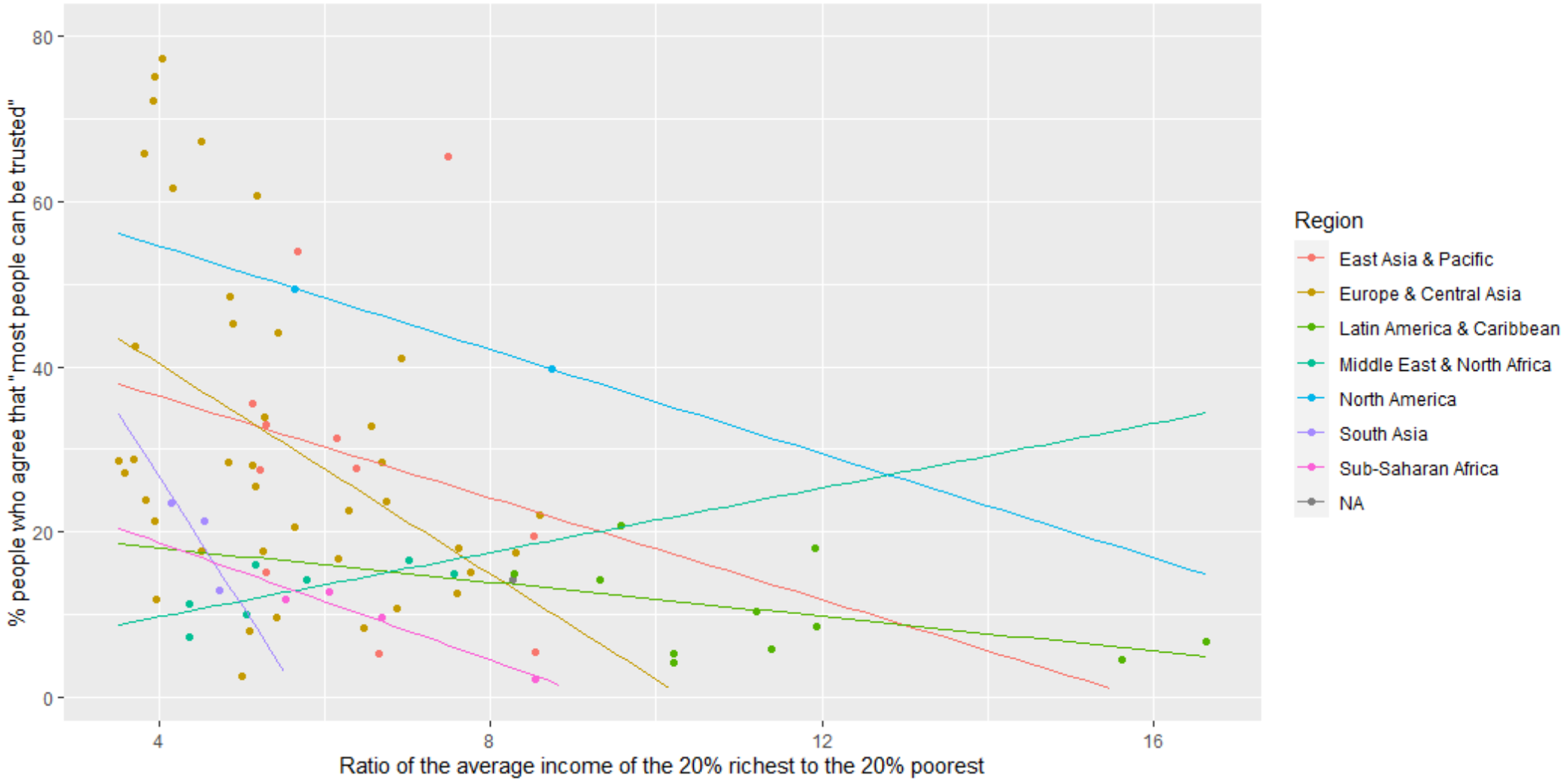
New data (WVS + World Bank)



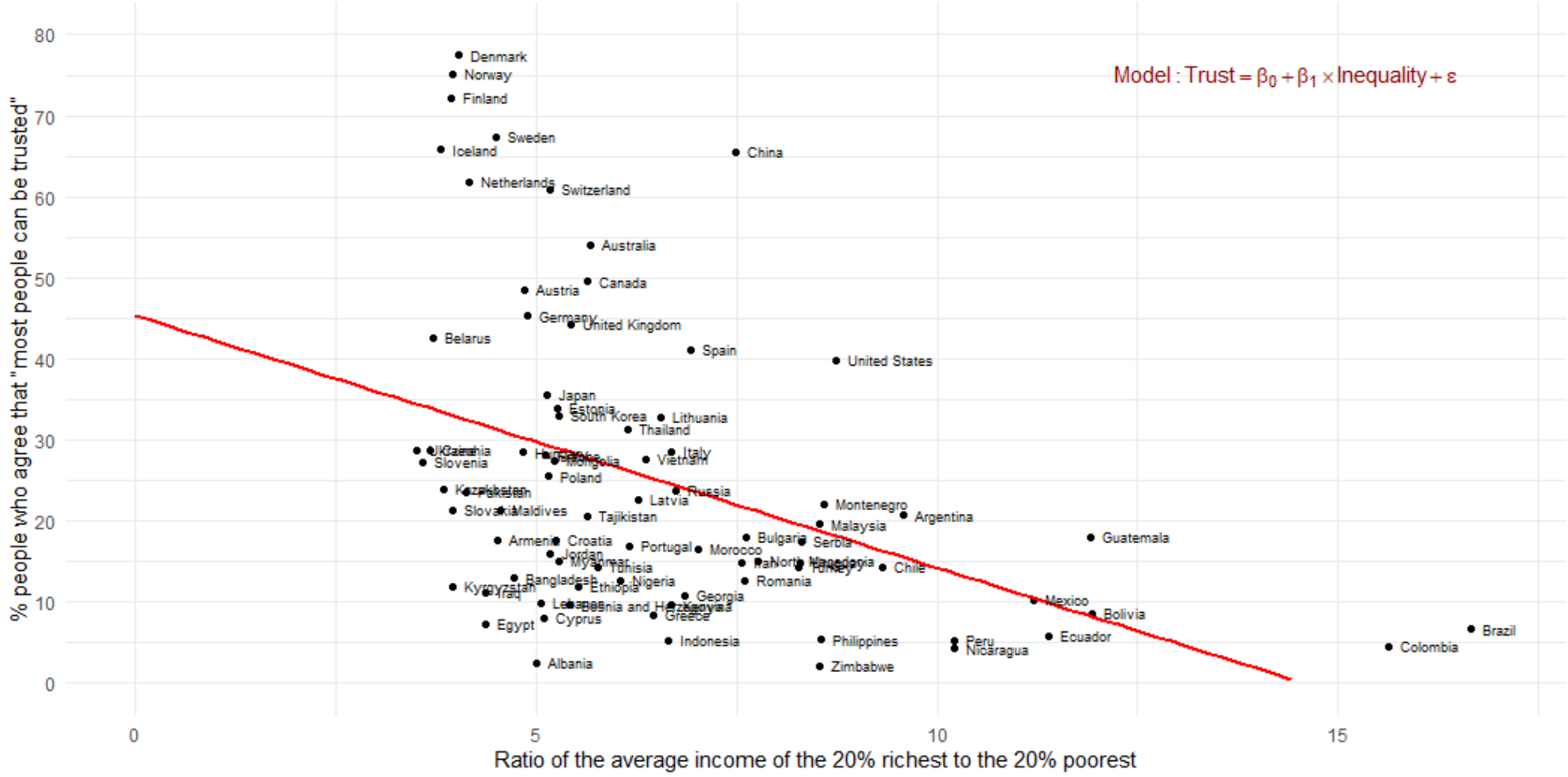
New data (WVS + World Bank)



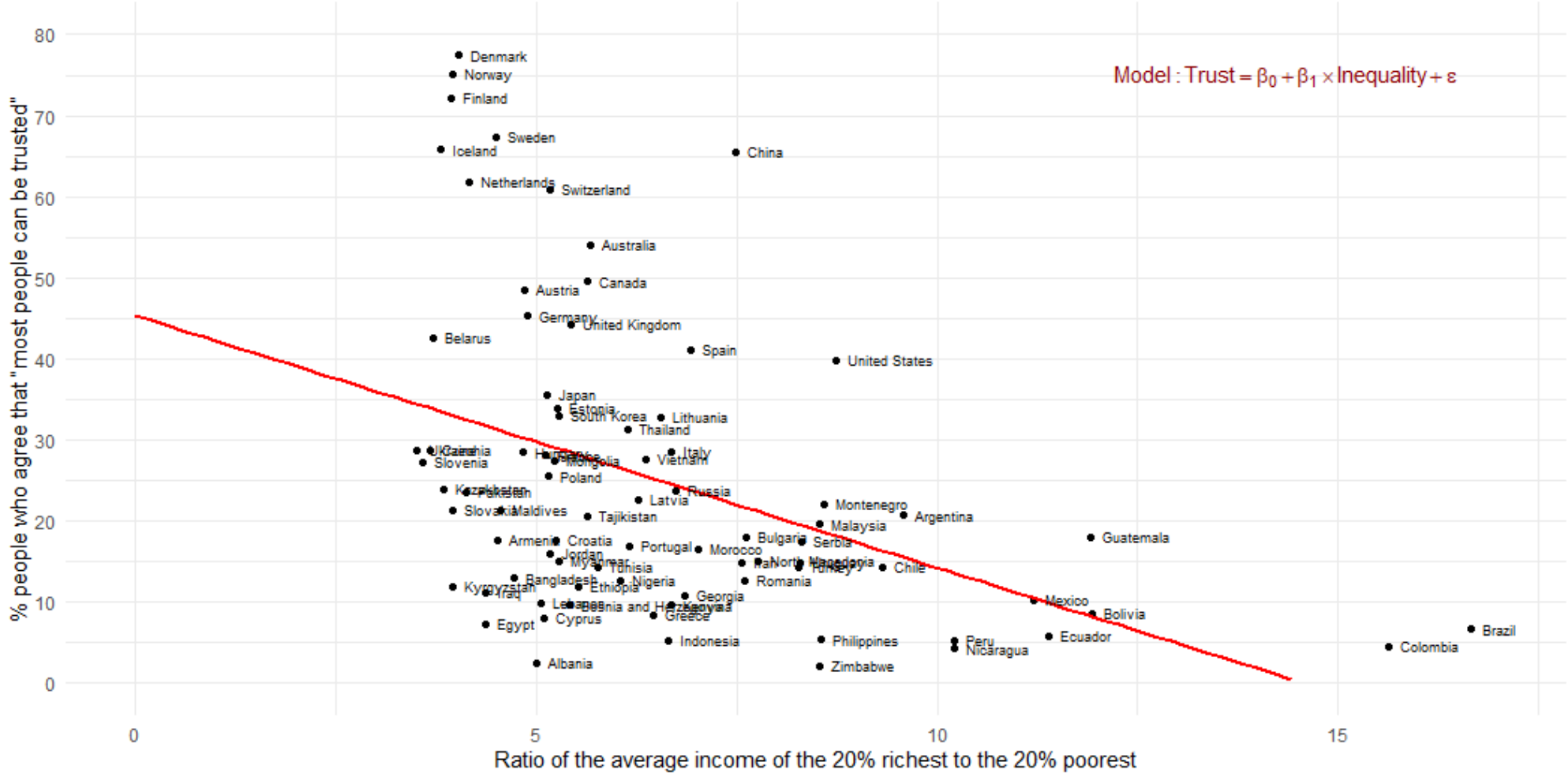
New data (WVS + World Bank)



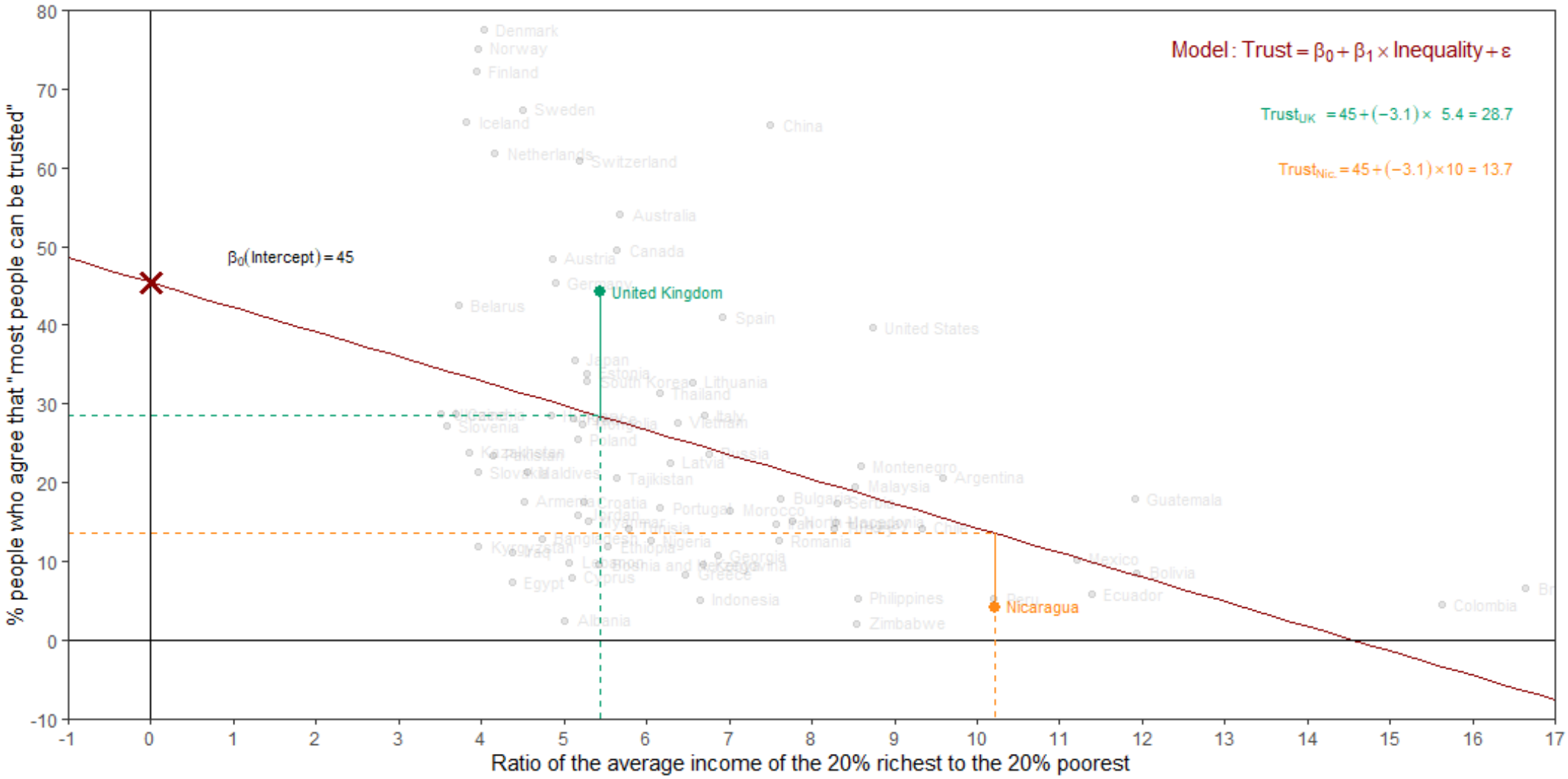
Model estimates



Model estimates

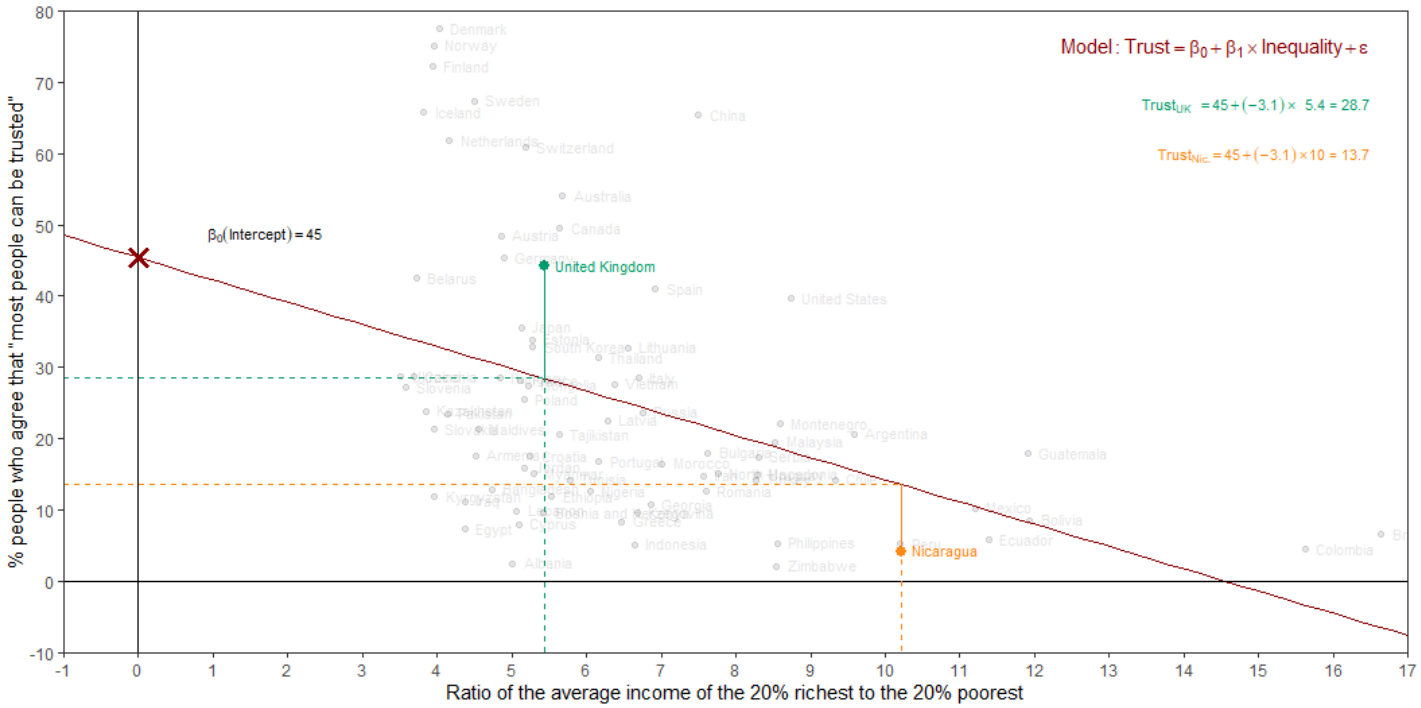


Model estimates



Model estimates

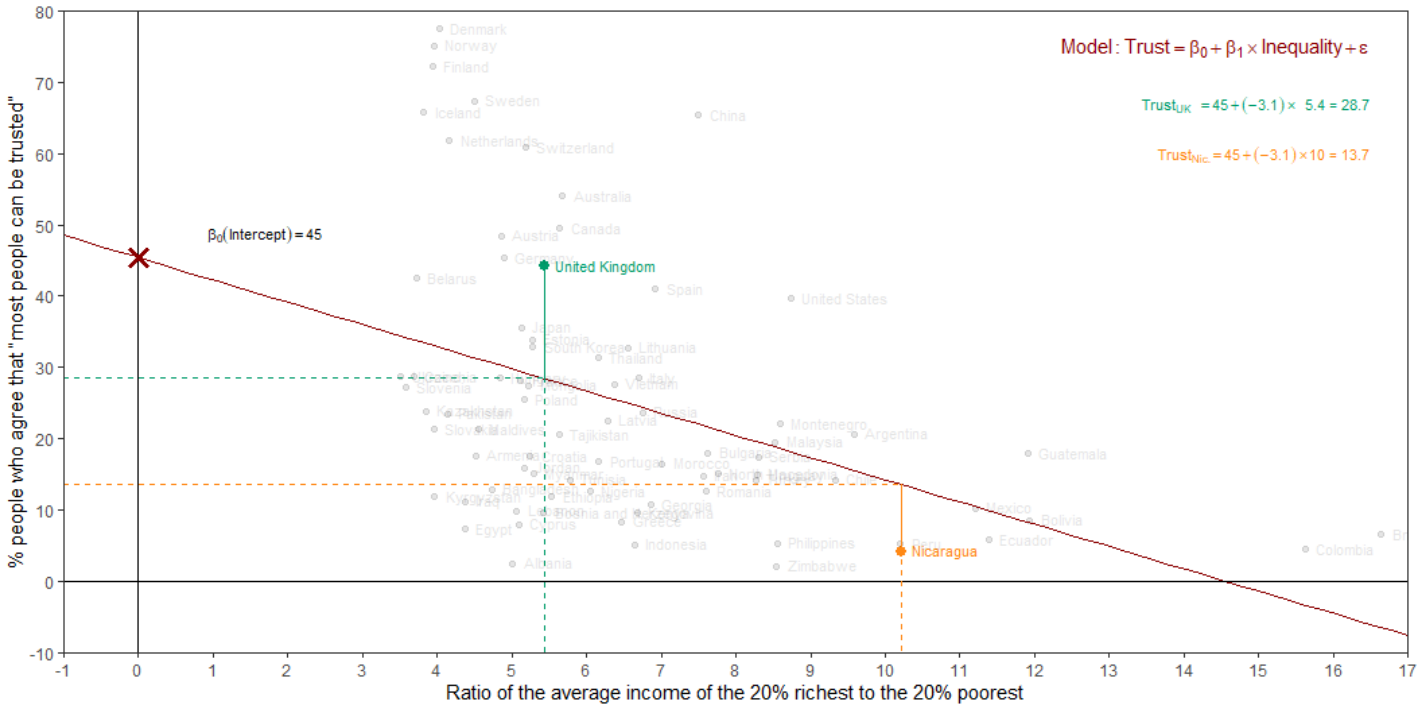
- **Intercept:** Social trust in countries with 0 inequality is expected to be 45.4 on average.
- **Slope:** For each additional point increase in Inequality, the model predicts the level of social trust to be lower, on average, by 3.1 points.



Model estimates

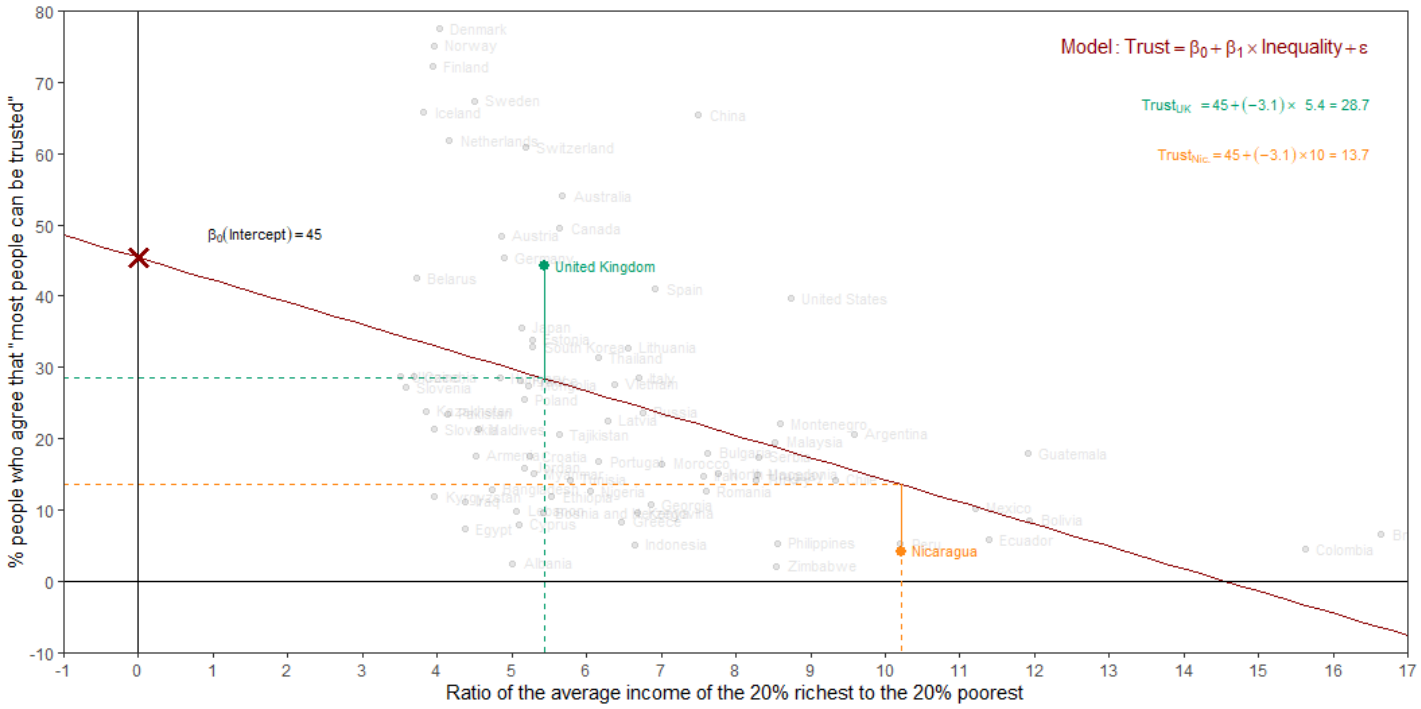
For each additional point increase in Inequality, the model predicts the level of social trust to be lower, on average, by 3.1 points

- This estimate is valid for the single sample of the countries in the model
- But what if we're not interested in quantifying the relationship between Inequality and Trust in only this sample?
- What if we want to say something about the relationship between these variables for all the countries in the world?



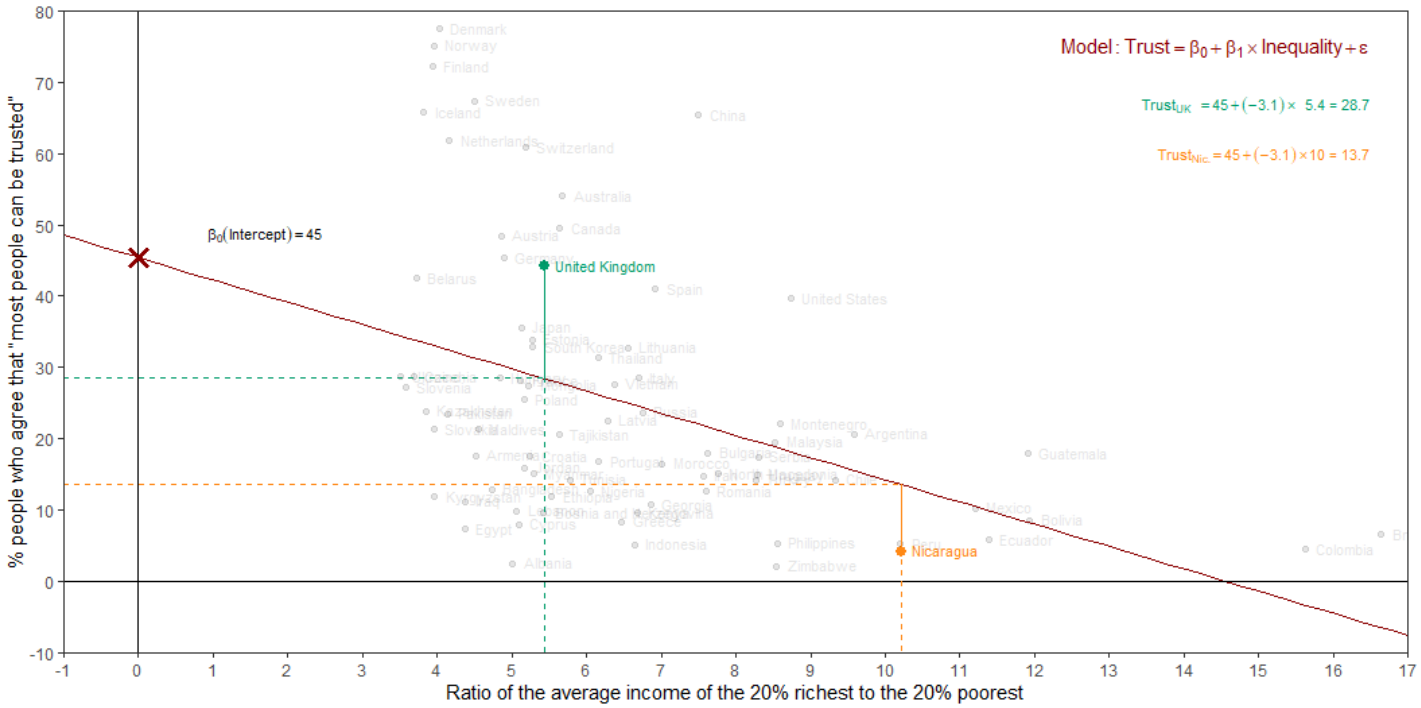
Hypothesis test for the slope

- "Do the data provide sufficient evidence that β_1) the true slope for the population) is different from 0?"
- **Null hypothesis** - $H_0: \beta_1 = 0$, there is no linear relationship between `inequality` and `trust`
- **Alternative hypothesis** - $H_A: \beta_1 \neq 0$,there is a linear relationship between `inequality` and `trust`



Hypothesis test for the slope

- Start with a null hypothesis, H_0 that represents the status quo
- Set an alternative hypothesis, H_A that represents the research question, i.e. what we're testing for
- Conduct a hypothesis test under the assumption that the null hypothesis is true and calculate a **p-value** (probability of observed or more extreme outcome given that the null hypothesis is true)
  - if the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis
  - if they do, then reject the null hypothesis in favour of the alternative



Application: Österman (2021)

Data from Österman (2021) "Can We Trust Education for Fostering Trust? Quasi-experimental Evidence on the Effect of Education and Tracking on Social Trust"

```
## Load the Osterman data:
osterman <- haven::read_dta("https://cgmoreh.github.io/SSC7001M/data/osterman.dta")
```

- cumulative European Social Survey (ESS) data, consisting of the nine rounds from 2002 to 2018
- data are weighted using ESS design weights (we will disregard this, sowe can expect our results to always differ somewhat!)
- follows 'the established approach of using a validated three-item scale' to study generalised social trust
- data also includes twenty-seven educational reforms implemented in sixteen European countries over six decades, where for each reform we can compare earlier reform-unaffected cohorts with later reform-affected cohorts

Österman, Marcus. 2021. 'Can We Trust Education for Fostering Trust? Quasi-Experimental Evidence on the Effect of Education and Tracking on Social Trust'. *Social Indicators Research* 154(1):211–33

(online)

Application: Österman (2021)

Data from Österman (2021) "Can We Trust Education for Fostering Trust? Quasi-experimental Evidence on the Effect of Education and Tracking on Social Trust"

- The scale consists of the classic trust question, an item on whether people try to be fair, and an item on whether people are helpful:
  - 'Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?'
  - 'Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair?'
  - 'Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves?'
- All of the items may be answered on a scale from 0 to 10 (where 10 represents the highest level of trust) and the scale is calculated as the mean of the three items
- The three-item scale improves measurement reliability and cross-country validity compared to using a single item, such as the classic trust question.

Österman, Marcus. 2021. 'Can We Trust Education for Fostering Trust? Quasi-Experimental Evidence on the Effect of Education and Tracking on Social Trust'. *Social Indicators Research* 154(1):211–33

(online)

Application: Österman (2021)

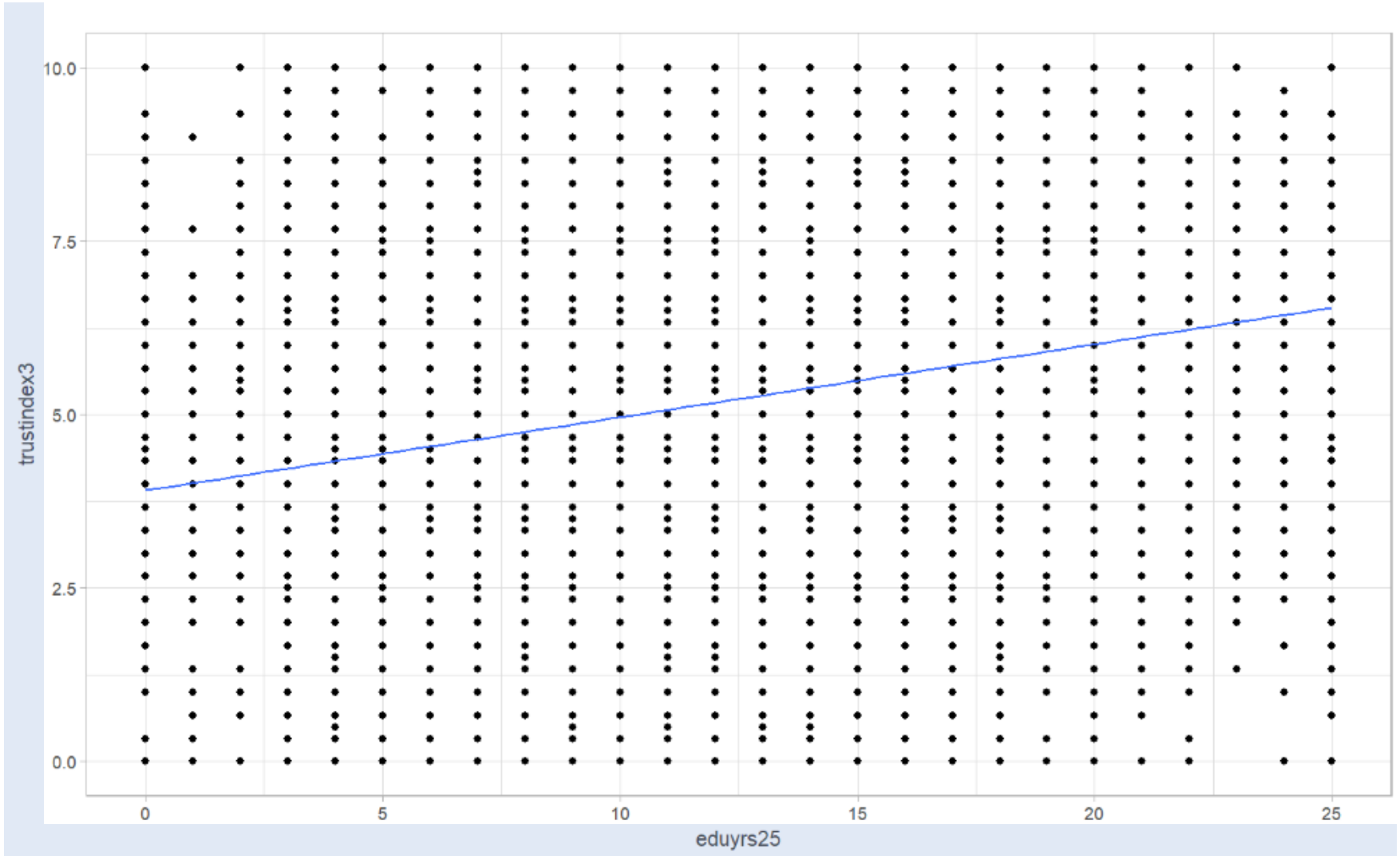
Table 2 in Österman (2021)

	N	Mean	SD	Min	Max
ppltrst	68733	5.097	2.395	0	10
pplfair	68548	5.729	2.212	0	10
pplhlp	68665	4.91	2.285	0	10
dscrgrp	68565	1.942	0.233	1	2
brncntr	68796	1	0	1	1
facntr	68796	0.961	0.194	0	1
mocntr	68796	0.964	0.185	0	1
yrbrn	68796	1959.058	11.914	1926	1993
agea	68796	50.949	12.244	25	80
female	68796	0.53	0.499	0	1
eduyrs25	68211	12.7	4.193	0	25
blgetmg_d	68796	0.015	0.122	0	1
ipudrst_rev	66903	7.317	2.063	0	10
fnotbrneur	68796	0.011	0.102	0	1
mnotbrneur	68796	0.009	0.096	0	1
trustindex3	68796	5.244	1.906	0	10
inst_trustindex5	59255	4.496	1.998	0	10
reform_id_num	68796	13.598	7.633	1	27
reform_years	68796	-0.174	4.305	-7	7
reform1_eonly_7	40632	0.5	0.5	0	1
reform1_d_7	29931	0.444	0.497	0	1
reform1_donly_7	8258	0.424	0.494	0	1
reform1_7	68796	0.482	0.5	0	1
paredu_a_high	64960	0.348	0.476	0	1
cntry_cohort	68796	1135.899	534.978	32	2003
fbrneur	68796	0.029	0.166	0	1
mbrneur	68796	0.026	0.16	0	1

Österman, Marcus. 2021. 'Can We Trust Education for Fostering Trust? Quasi-Experimental Evidence on the Effect of Education and Tracking on Social Trust'. *Social Indicators Research* 154(1):211–33

([online])

- A scatter-plot of 'trustindex3' by 'eduyrs25'



Does education predict trust?

```
m_edu <- lm(trustindex3 ~ eduyrs25, data = osterman)
summary(m_edu)
```

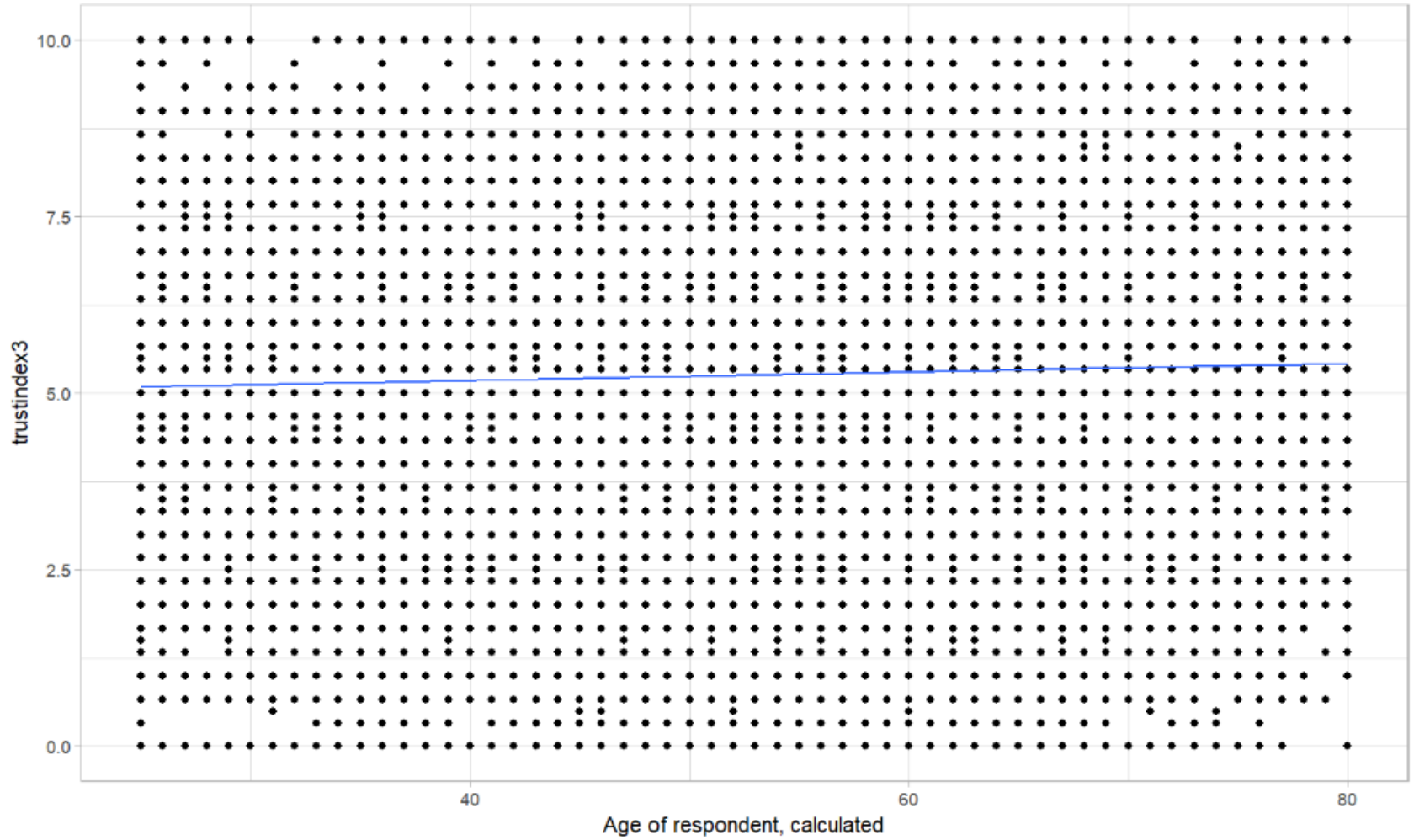
```
##
## Call:
## lm(formula = trustindex3 ~ eduyrs25, data = osterman)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.544 -1.174   0.143   1.299   6.091
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.90859    0.02263   172.7   <2e-16 ***
## eduyrs25     0.10542    0.00169    62.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.85 on 68209 degrees of freedom
##   (585 observations deleted due to missingness)
## Multiple R-squared:  0.0538,    Adjusted R-squared:  0.0538
## F-statistic: 3.88e+03 on 1 and 68209 DF,  p-value: <2e-16
```

The "slope" coefficient; rather weak, but more years of education completed is associated with higher levels of social trust

p-value; very small; smaller than 0.001, and definitely smaller than 0.05 (for what it's worth)

Conclusion:
One's education appears to have an impact on their level of social trust. Each additional year of completed formal education is associated with a 0.105-point increase on the measured "social trust" scale. Based on the p-value (<0.001), we reject the null-hypothesis that this relationship would appear in our data by chance only, and provisionally accept that education affects social trust in the wider EU population.

Does age predict trust?



```
m_age <- lm(trustindex3 ~ agea, data = osterman)
summary(m_age)
```

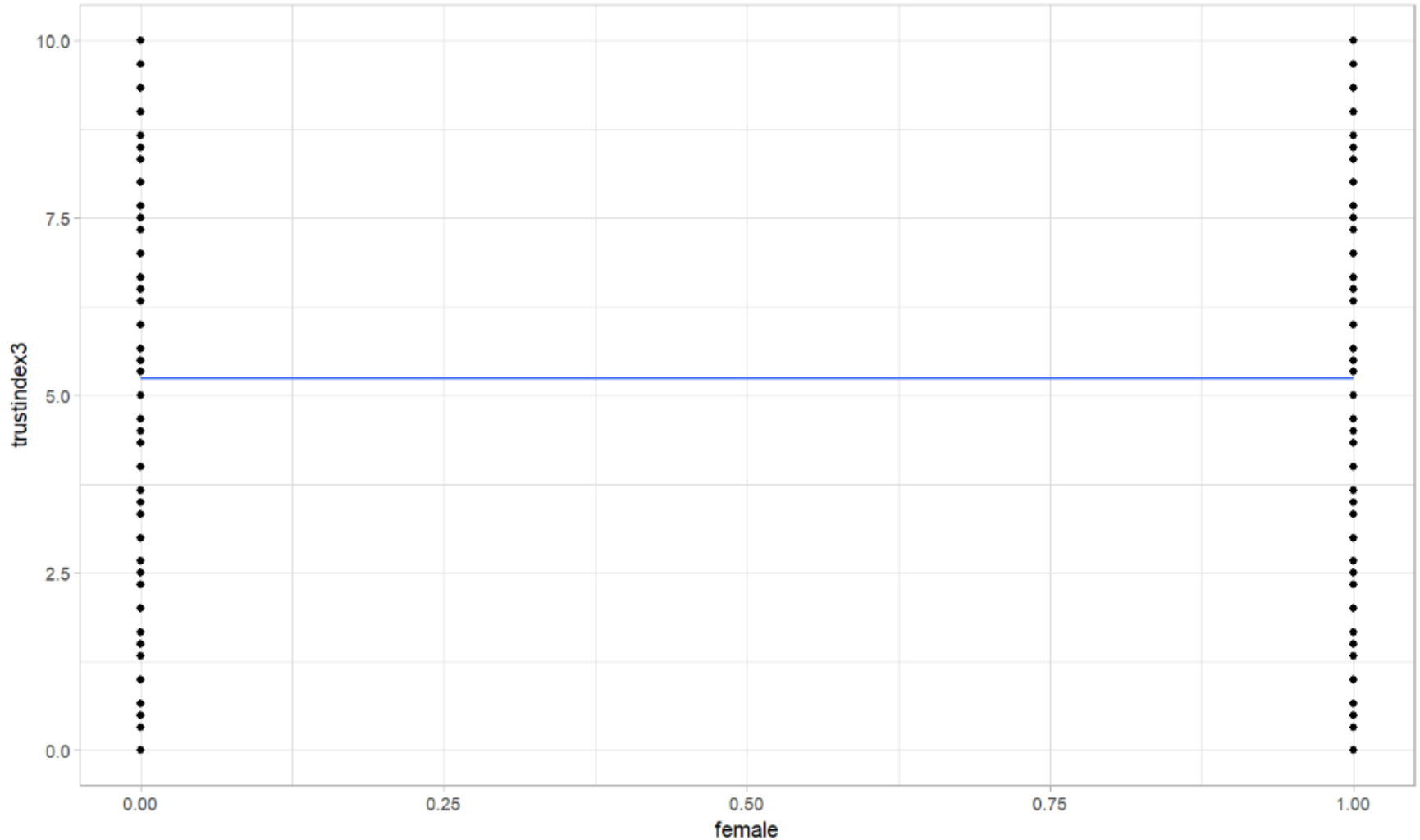
```
##
## Call:
## lm(formula = trustindex3 ~ agea, data = osterman)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.418 -1.244  0.119  1.405  4.913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.936953   0.031078   158.9   <2e-16 ***
## agea        0.006019   0.000593    10.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.9 on 68794 degrees of freedom
## Multiple R-squared:  0.00149,    Adjusted R-squared:  0.00148
## F-statistic:  103 on 1 and 68794 DF,  p-value: <2e-16
```

The "slope" coefficient; rather weak, but being older is associated with higher levels of social trust

p-value; very small; smaller than 0.001, and definitely smaller than 0.05 (for what it's worth)

Conclusion:
One's age appears to have an impact on their level of social trust. This association is weak in substantive terms. Each additional year of age is associated with a 0.006-point increase on the measured "social trust" scale. Based on the p-value (<0.001), we reject the null-hypothesis that this relationship would appear in our data by chance only, and provisionally accept that age affects social trust in the wider EU population.

Does gender predict trust?



Does gender predict trust?

```
m_gender <- lm(trustindex3 ~ female, data = osterman)
summary(m_gender)
```

```
##
## Call:
## lm(formula = trustindex3 ~ female, data = osterman)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.247 -1.247   0.094   1.419   4.761
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.2393     0.0106  494.51   <2e-16 ***
## female        0.0081     0.0146    0.56     0.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.91 on 68794 degrees of freedom
## Multiple R-squared:  4.5e-06,    Adjusted R-squared:  -1e-05
## F-statistic: 0.309 on 1 and 68794 DF,  p-value: 0.578
```

The "slope" coefficient; very weak, but being a woman is associated with higher levels of social trust compared to being a man

p-value; it runs on a 0-1 scale, so 0.58 (i.e. 58%) is very high, and much much higher than 0.05 (for what it's worth)

Conclusion:
One's sex doesn't appear to have an impact on their level of social trust. The average score of women on the social trust scale is merely 0.008 points higher than that of men. Based on the p-value (0.58), we cannot reject the null-hypothesis that the difference between men and women in the wider EU population is in fact equal to 0 (as opposed to 0.008). There is in fact a probability of 0.58 (or 58%) that we would be wrong in rejecting the null-hypothesis. In conclusion, we don't have evidence to suggest that sex is associated with differences in levels of social trust.

Multiple linear regression

Multiple linear regression

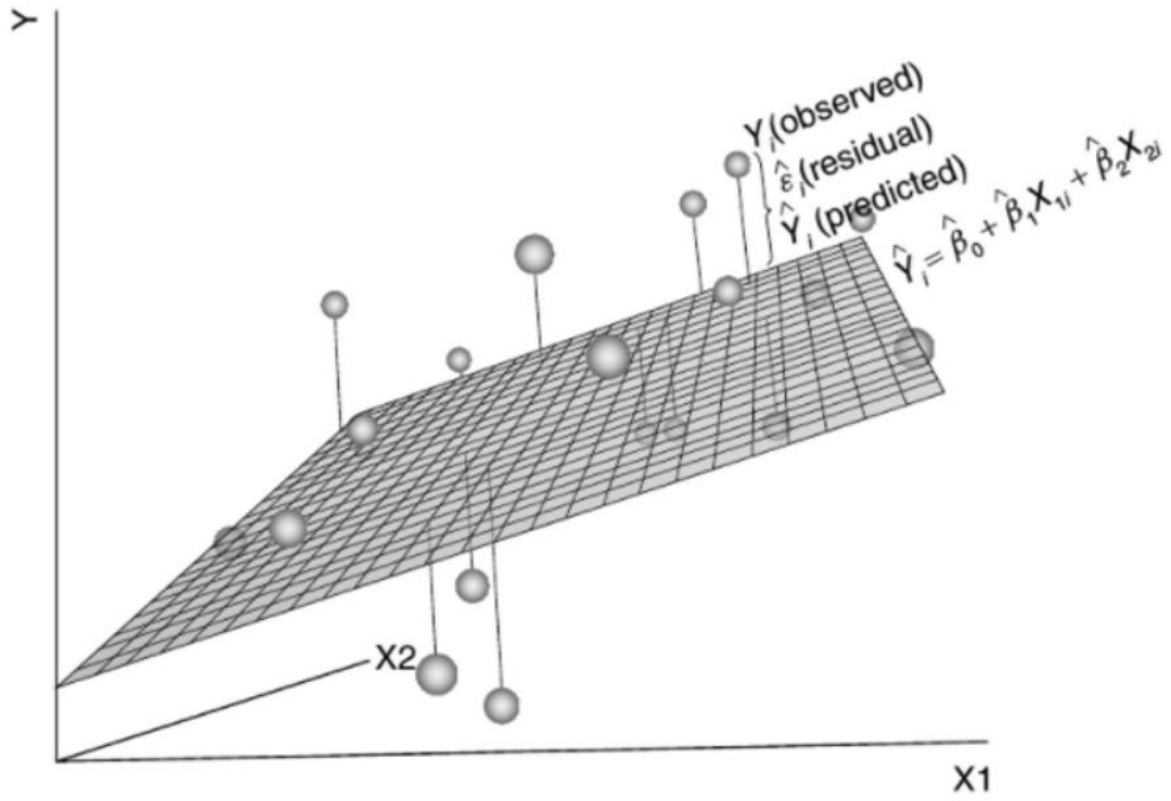
- it allows us to test the joint effects of more than one predictor
- usually, we are interested in one primary predictor, but *while taking into account*, or *holding constant* the values of other predictors
- it allows us to account for the possible effect of other factors (variables) that may influence our regression results
- the multiple regression model for p number of predictors can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + e$$

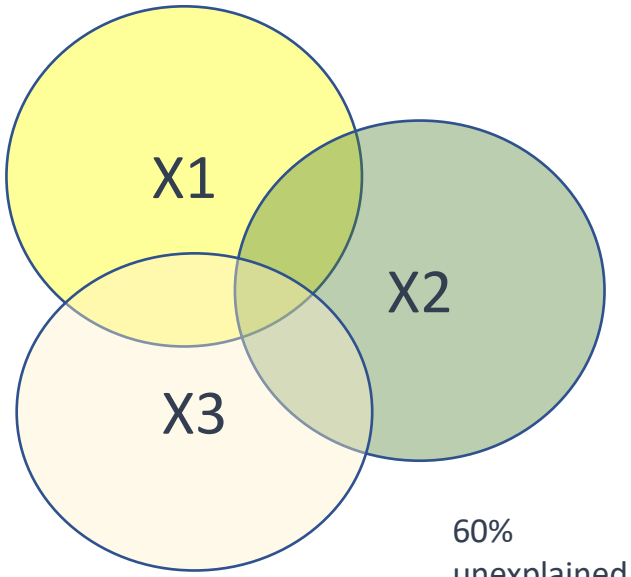
- but visualising it is a bit more difficult...

Multiple linear regression

Multiple linear regression



$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \hat{\beta}_p \bar{X}_p$$



Does education + age + gender predict/explain trust?

```
##
## Call:
## lm(formula = trustindex3 ~ eduyrs25 + agea + female, data = osterman)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.746 -1.176  0.133  1.312  6.071
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.954729   0.042894   68.88  <2e-16 ***
## eduyrs25    0.116400   0.001735   67.11  <2e-16 ***
## agea        0.015558   0.000594   26.20  <2e-16 ***
## female      0.041157   0.014151    2.91  0.0036 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.84 on 68207 degrees of freedom
##   (585 observations deleted due to missingness)
## Multiple R-squared:  0.0634,    Adjusted R-squared:  0.0633
## F-statistic: 1.54e+03 on 3 and 68207 DF,  p-value: <2e-16
```

We are now interpreting each "slope" coefficient, keeping in mind that they represent the effect of that variable *while keeping constant (eliminating) the effect of the other variables in the model*

p-value; small!; smaller than 0.01; note that this has changed compared to the previous model; now, once we also account for education and age (i.e. keeping the effect of those variables constant, one's sex seems to have an impact on their level of social trust

This is now the p-value for the entire model

Conclusion:

In a statistical model that predicts levels of social trust from education, age and gender, we find that all three predictors are associated with social trust. Having more education, being older and being a woman is associated with higher levels of social trust.