PSC 202

SYRACUSE UNIVERSITY

INTRODUCTION TO POLITICAL ANALYSIS

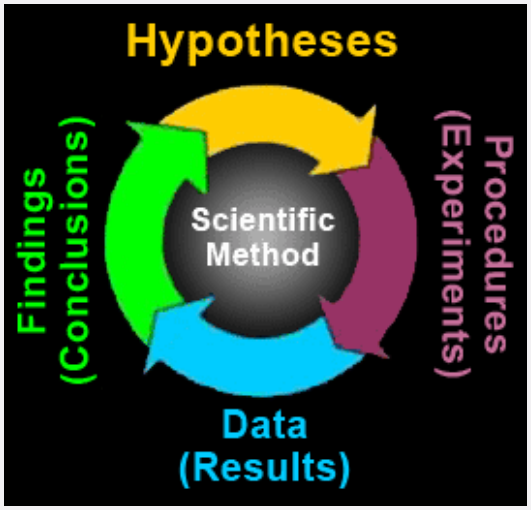
BIVARIATE HYPOTHESIS TESTING PART 2

EXAM

- Next week Monday: Exam #2
- Wednesday: Review
  - Please email questions etc. by tomorrow evening

WHERE WE ARE

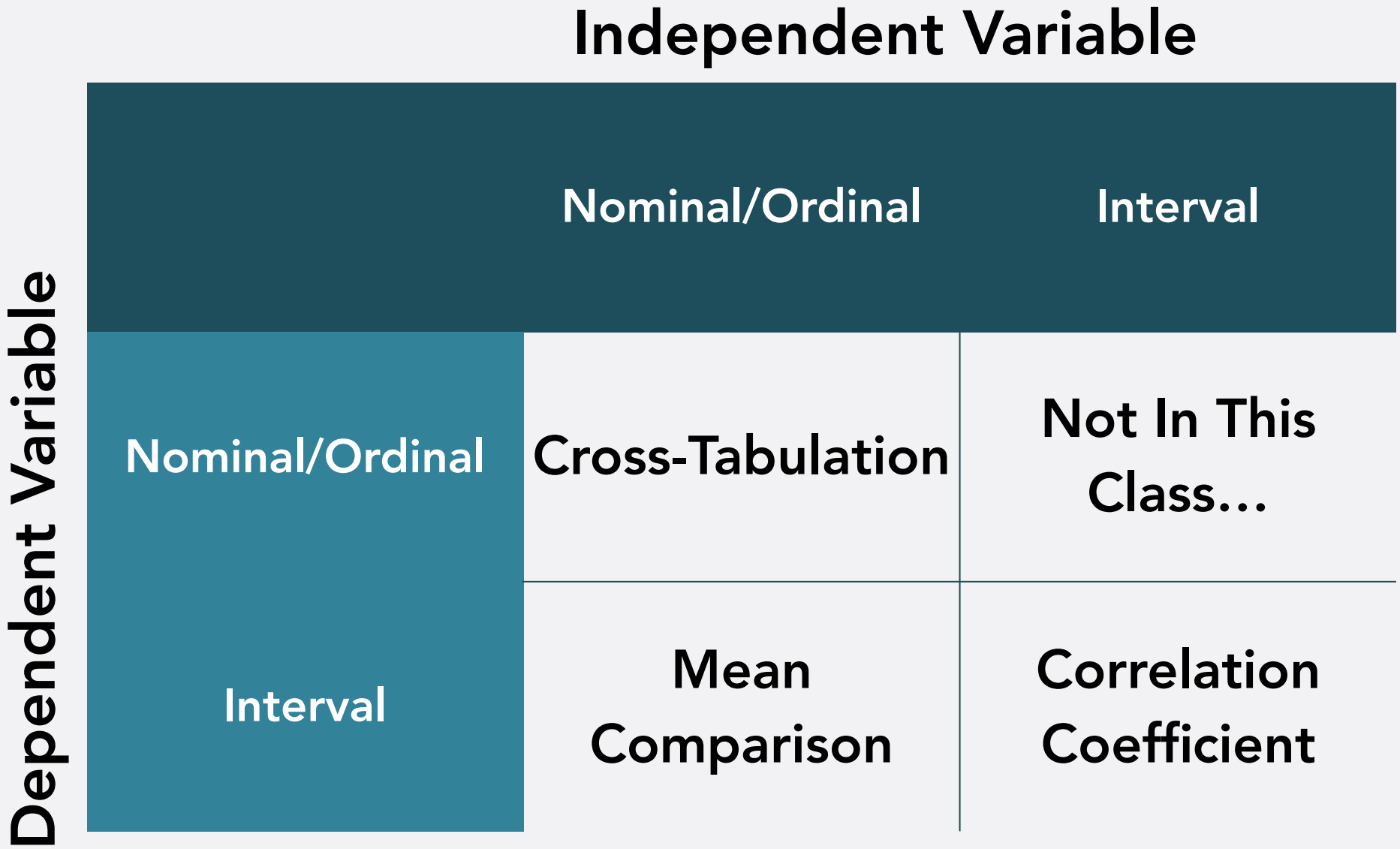
- Formulate research question
- Propose explanation/theory, hypotheses
- Data collection process
- Use data to evaluate hypotheses
- Reassess explanation



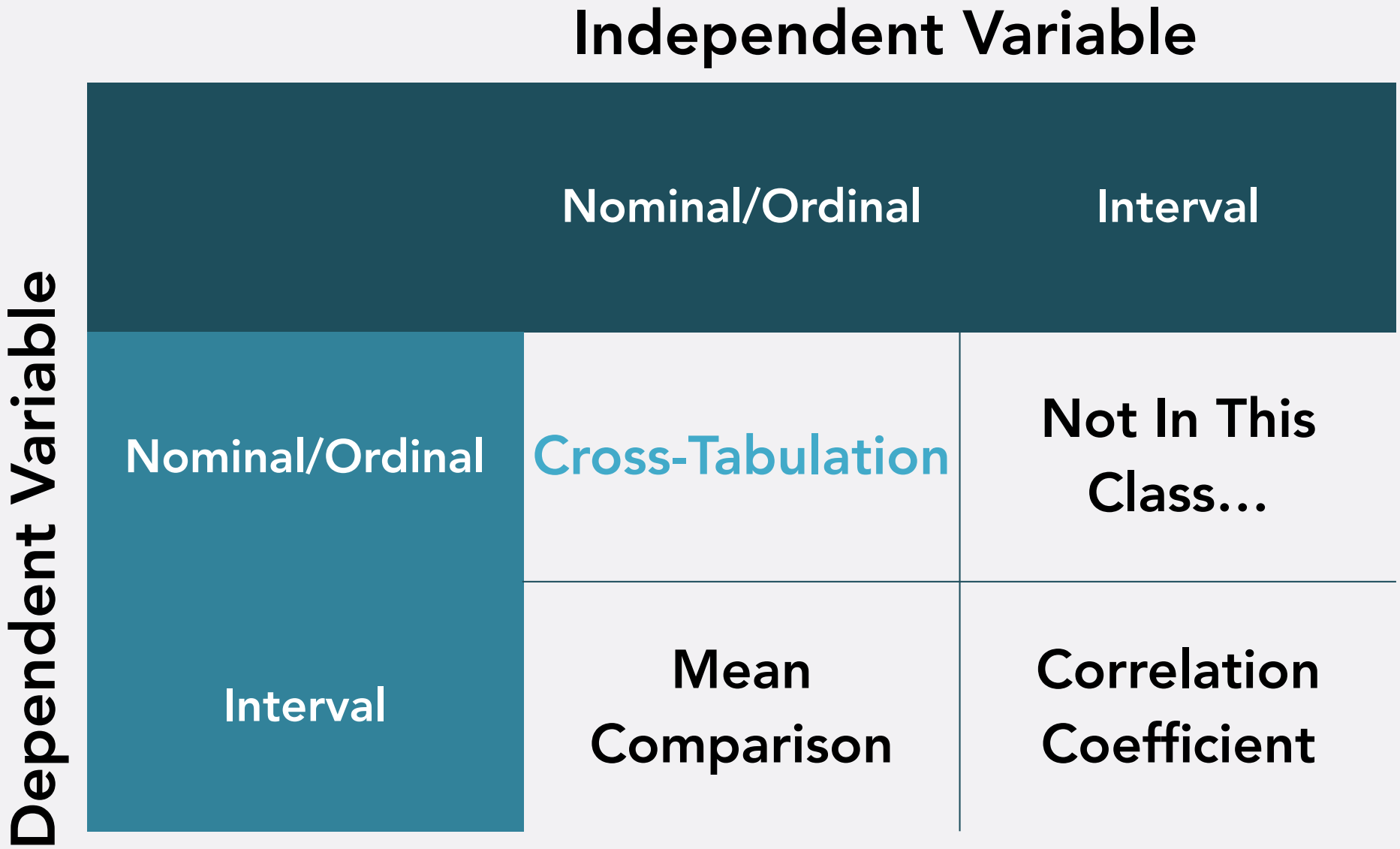
HURDLES

- Is there a credible causal mechanism that connects X to Y?
- Can we rule out the possibility that Y could cause X?
- Is there covariation between X and Y?
- Have we controlled for all confounding variables (Z) that might make the association between X and Y spurious?

BIVARIATE RELATIONSHIPS



BIVARIATE RELATIONSHIPS



CROSS-TABULATIONS

Gender

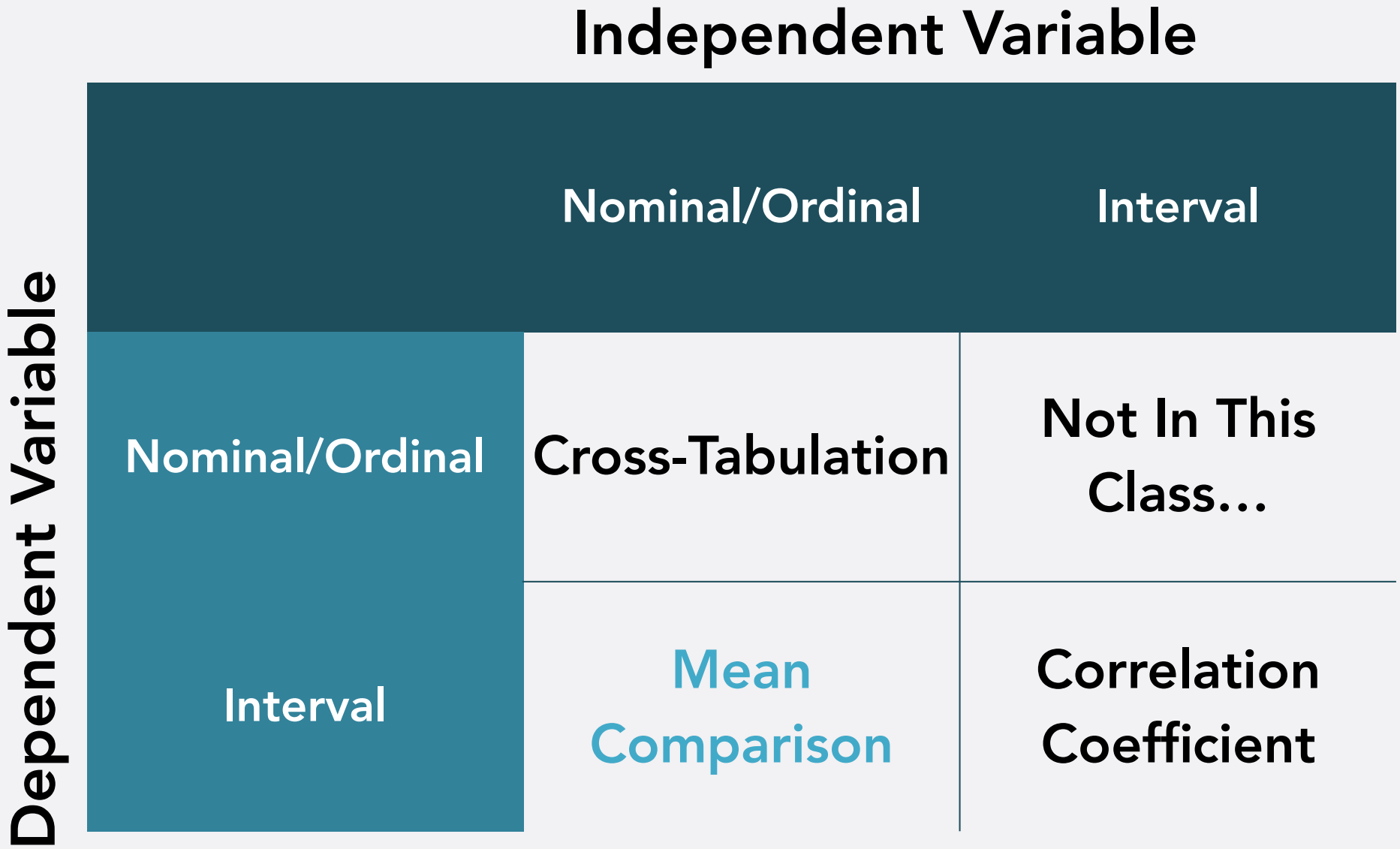
Approve of Biden	Male	Female	Total
Approve	44.0% (11)	52.4% (22)	49.2% (33)
Do Not Approve	56.0% (14)	47.6% (20)	50.8% (34)
Total	100% (25)	100% (42)	100% (67)

CROSS-TABULATIONS

Gender

Approve of Biden	Male	Female	Total
		8.4%	
Approve	44.0% (11)	52.4% (22)	49.2% (33)
Do Not Approve	56.0% (14)	47.6% (20)	50.8% (34)
Total	100% (25)	100% (42)	100% (67)

BIVARIATE RELATIONSHIPS



DEMOCRATIC PARTY

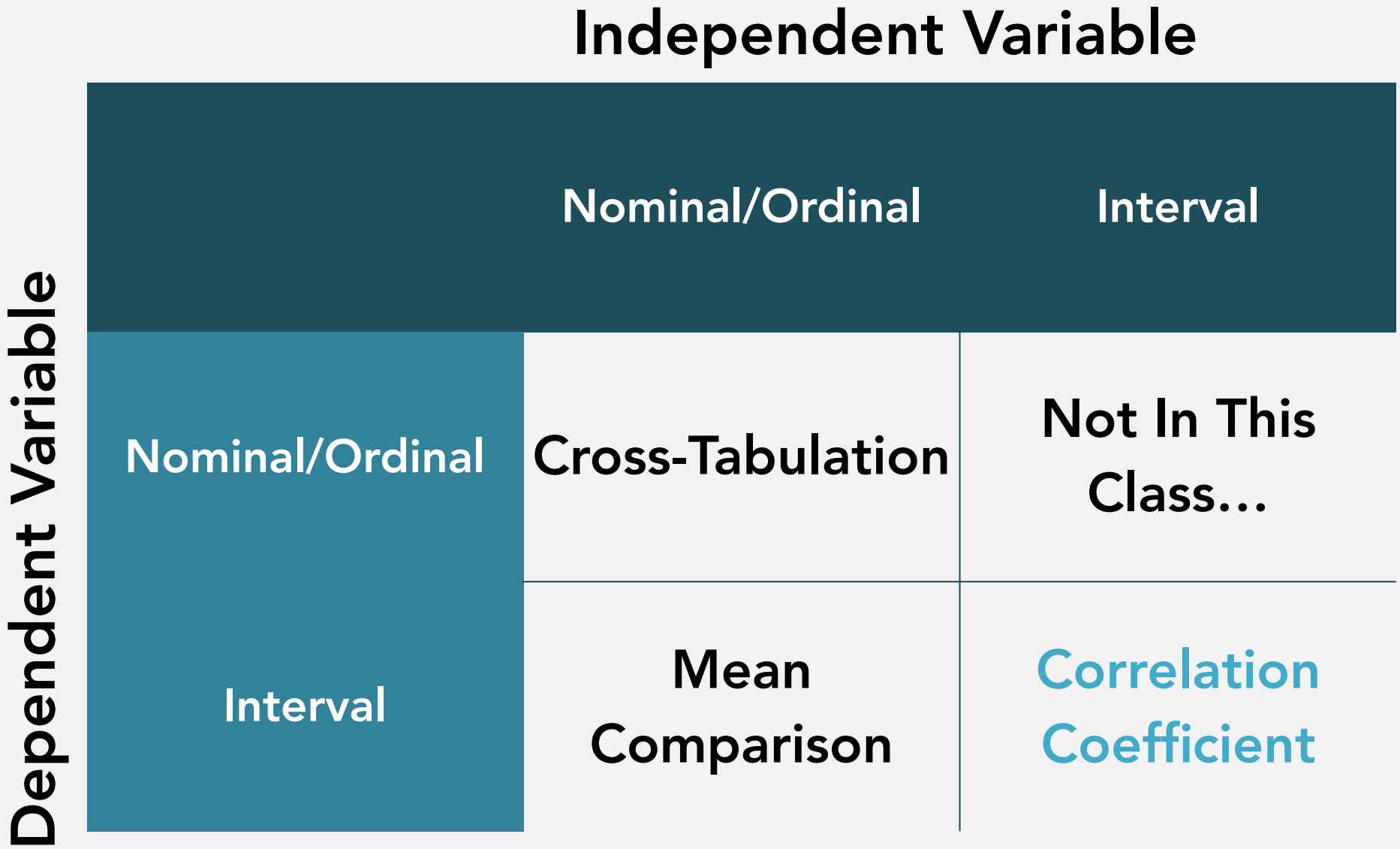
	Mean Thermometer Score	Frequency
Female	57.9	54
Male	50.0	27
Total	55.6	81

ZERO-ORDER RELATIONSHIP

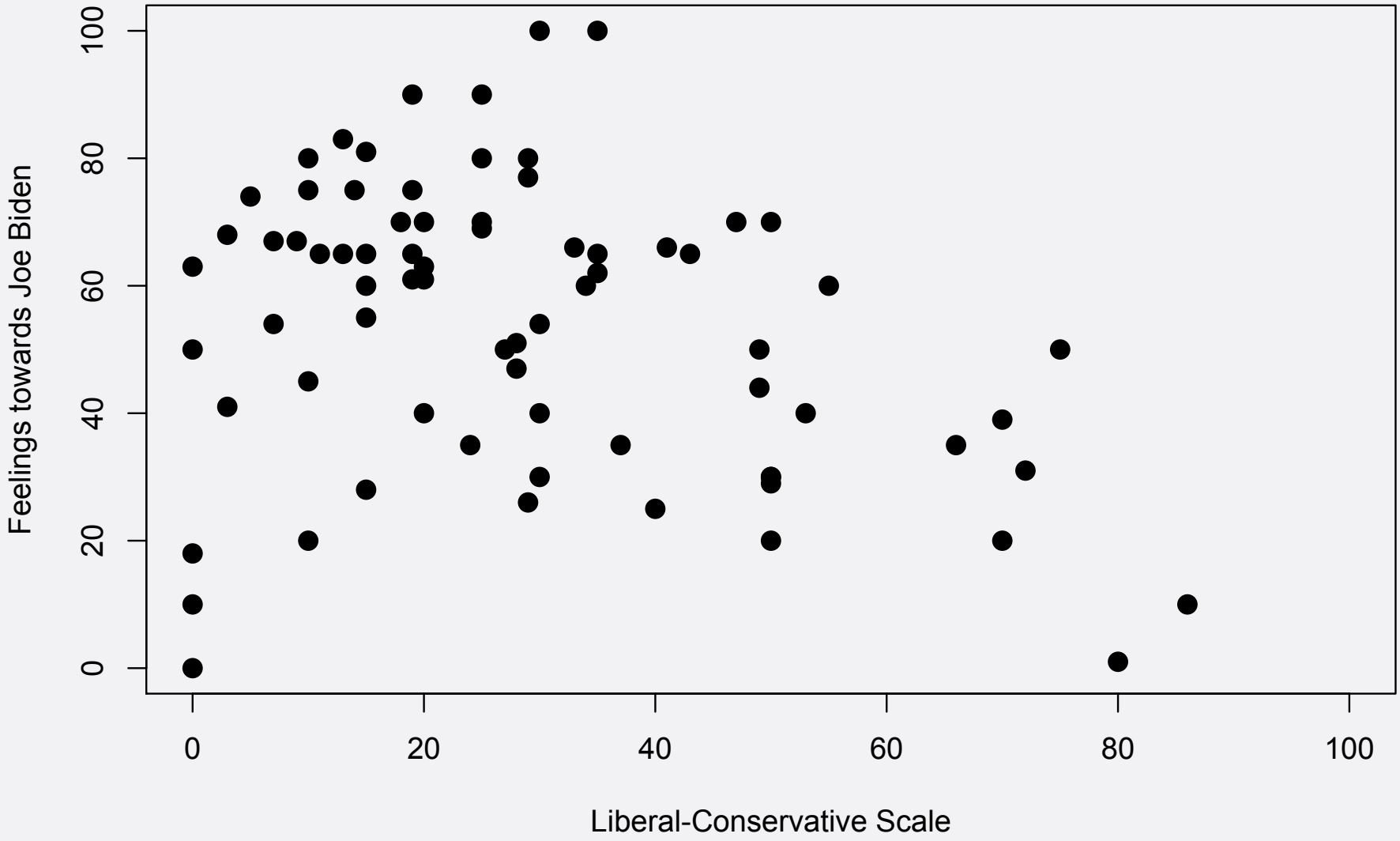
	Mean Thermometer Score	Frequency
Female	57.9	54
Male	50.0	27
Total	55.6	81

7.9

BIVARIATE RELATIONSHIPS

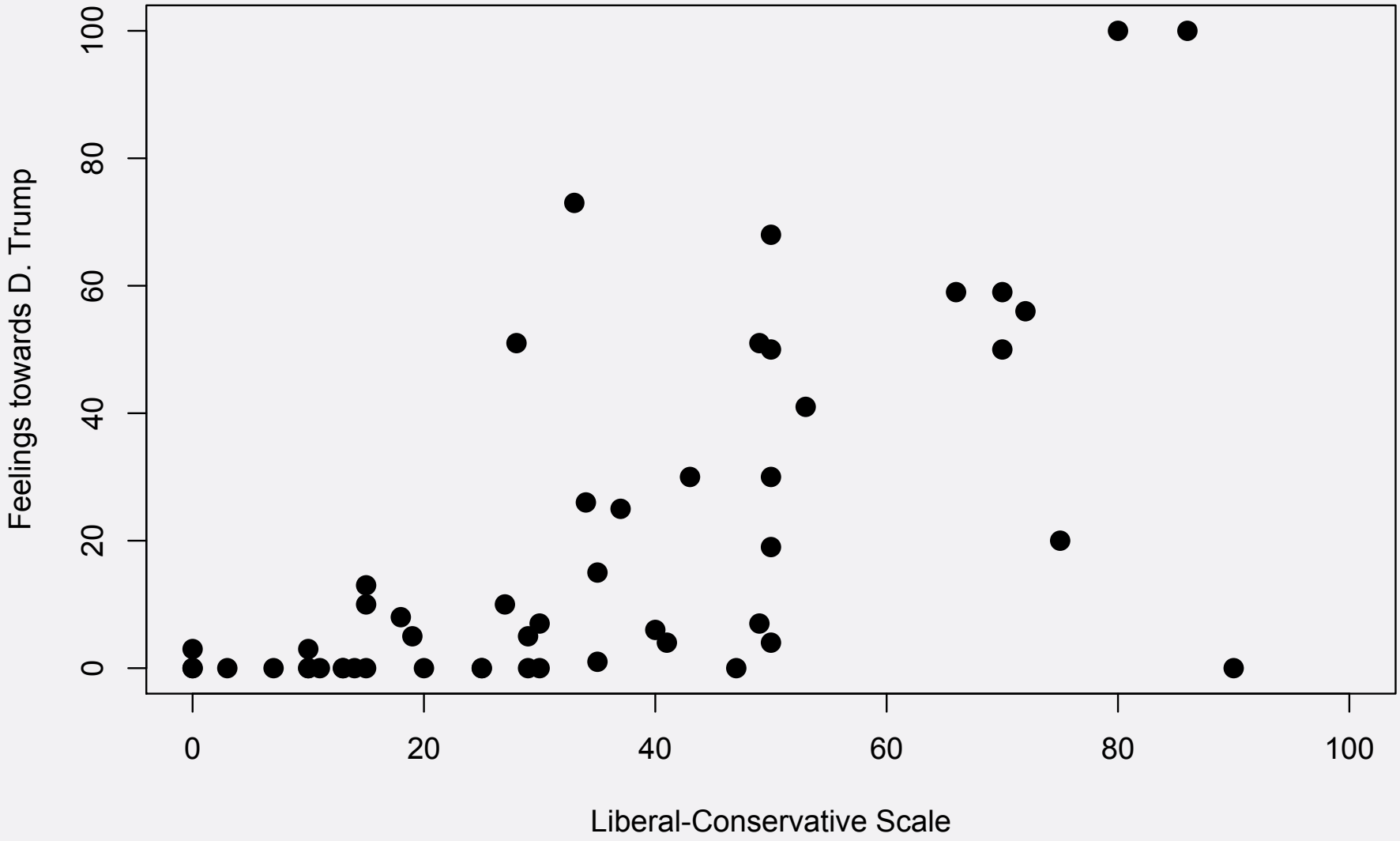


JOE BIDEN



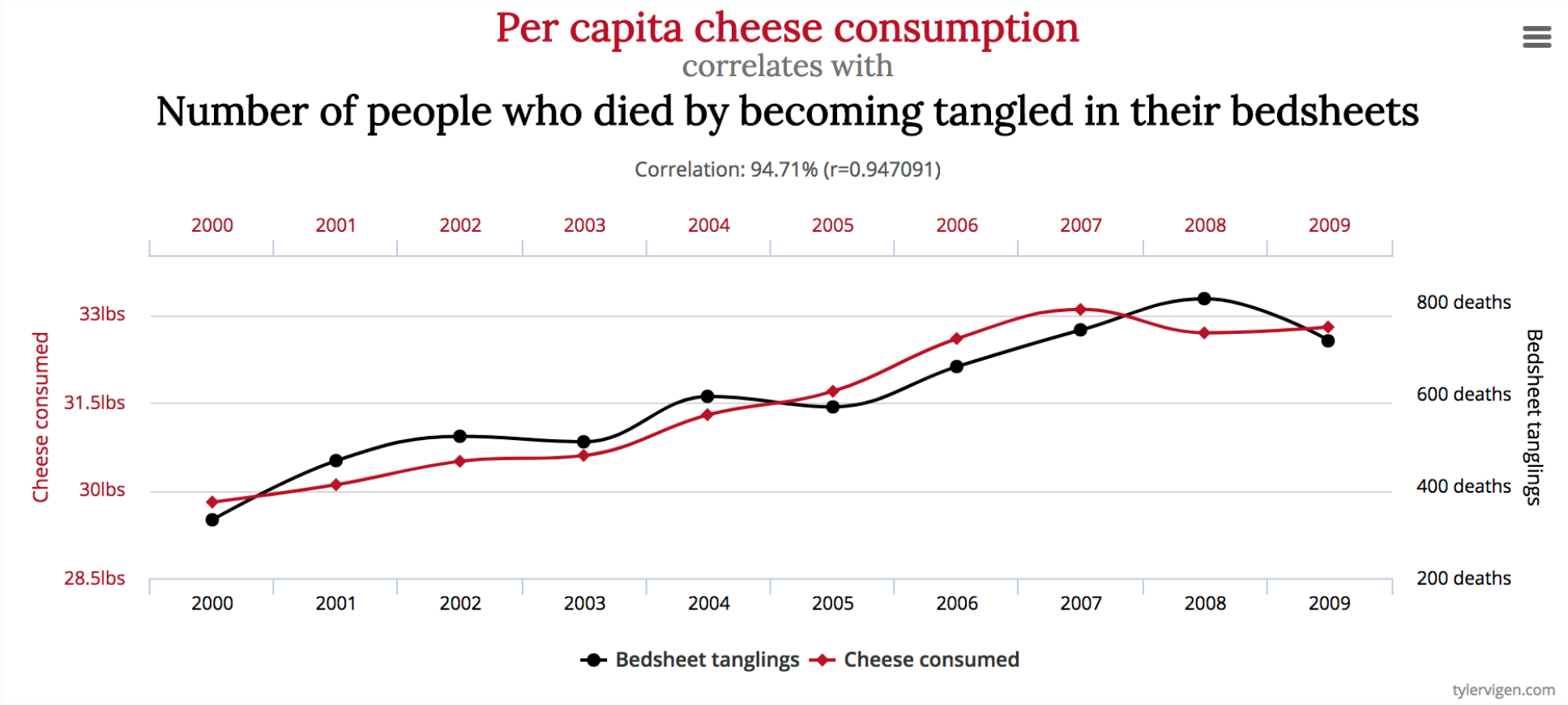
r=-0.32

DONALD TRUMP



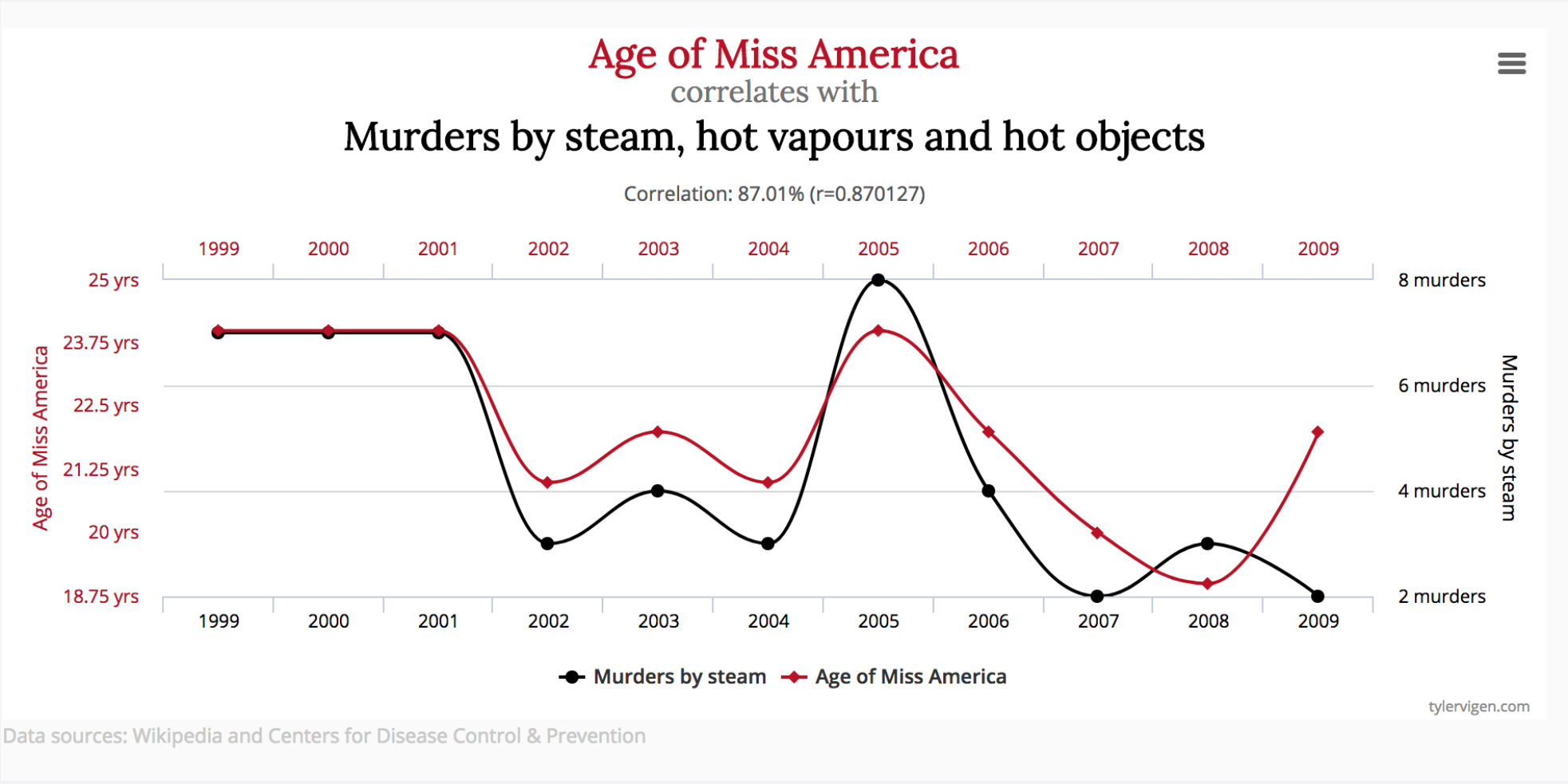
r=0.67

CAREFUL!



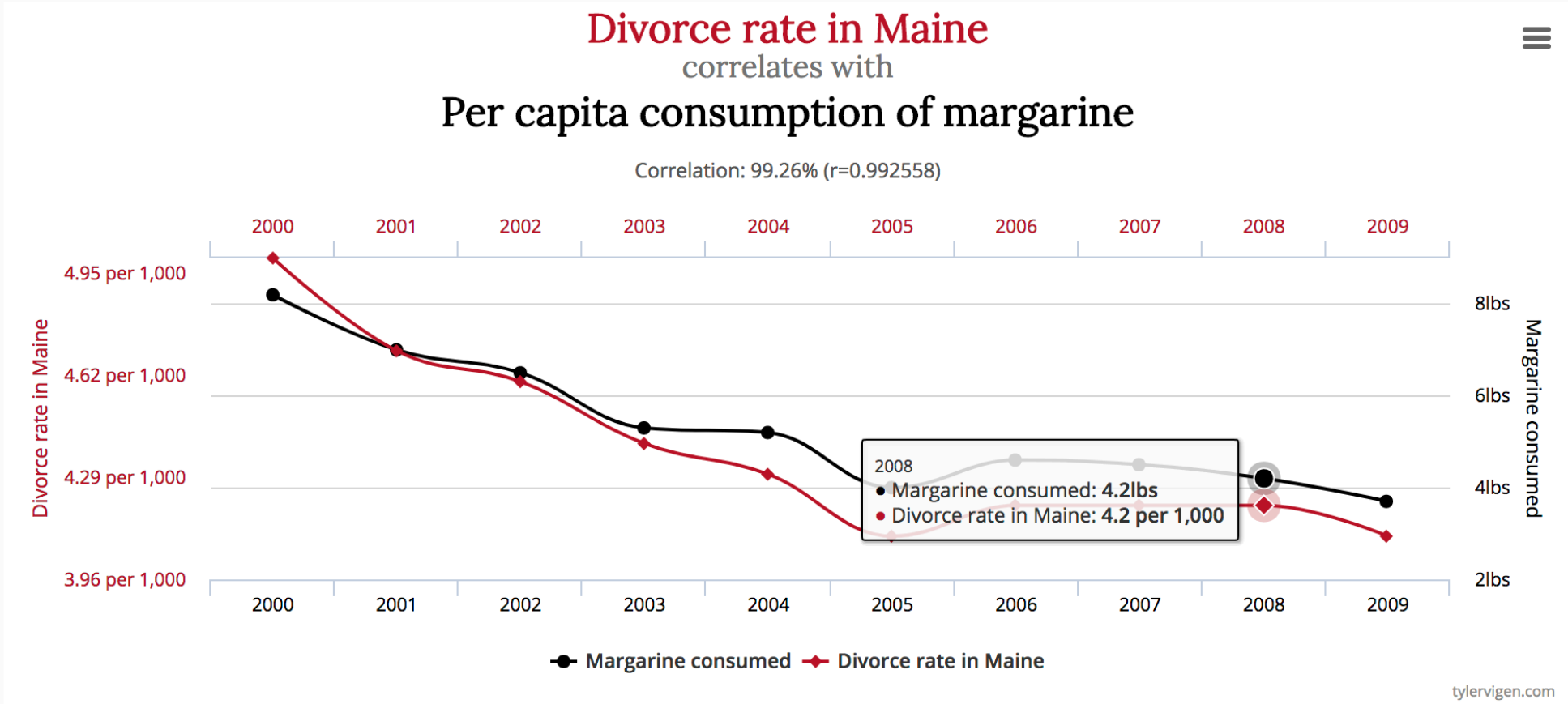
Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

CAREFUL!



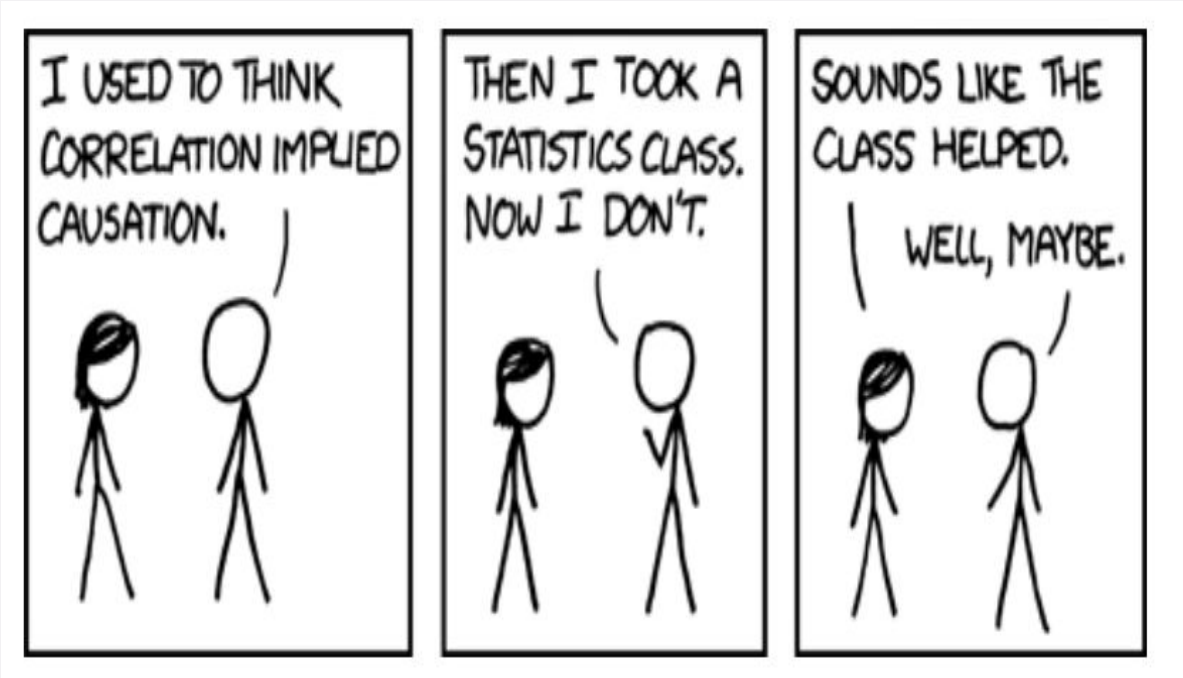
Data sources: Wikipedia and Centers for Disease Control & Prevention

CAREFUL!



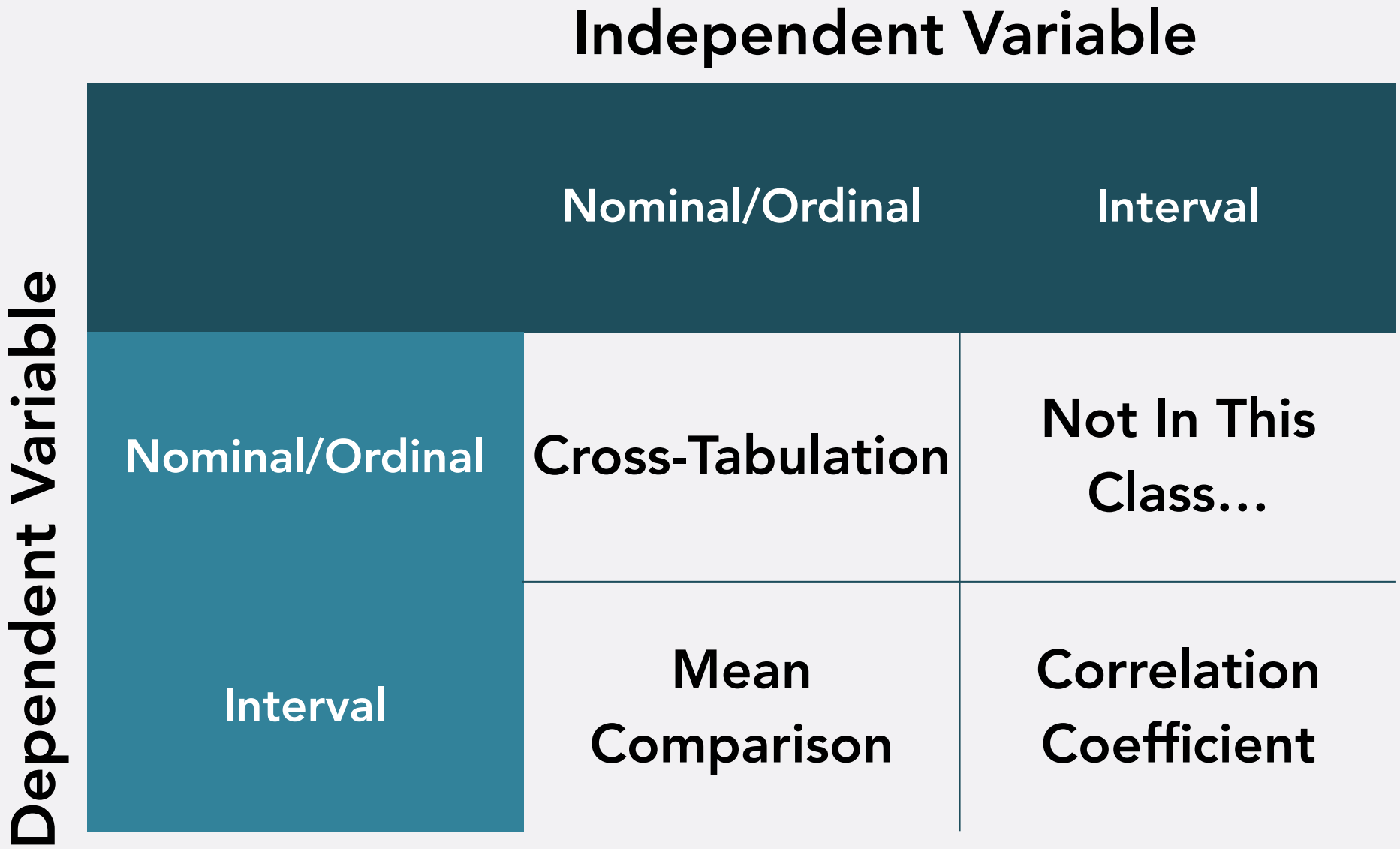
Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

CAREFUL!

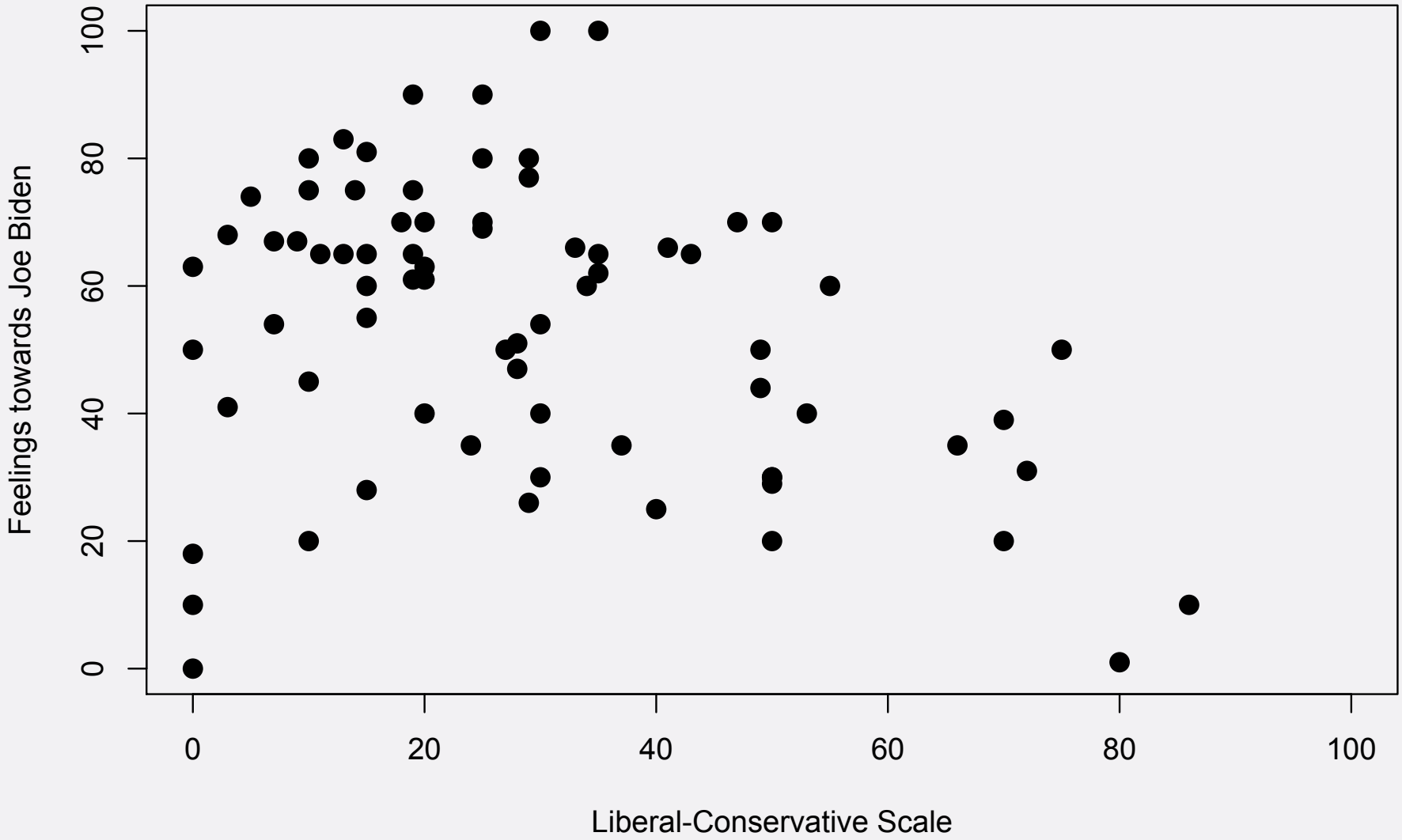


- **Important: Just because we find a correlation between two variables does **not** mean that the independent variable *causes* the dependent variable**
  - **The other hurdles to causality still apply!**

BIVARIATE RELATIONSHIPS

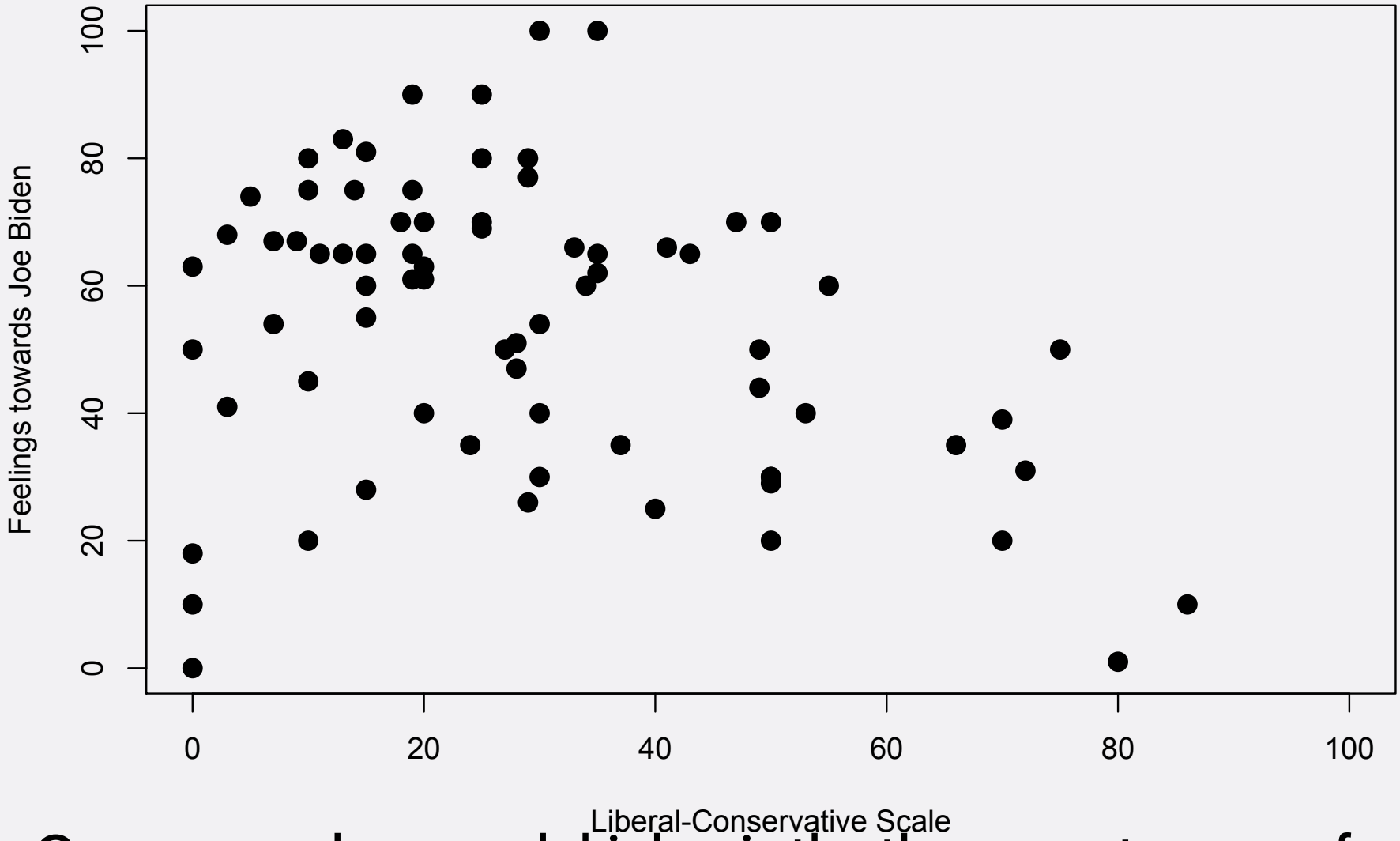


JOE BIDEN



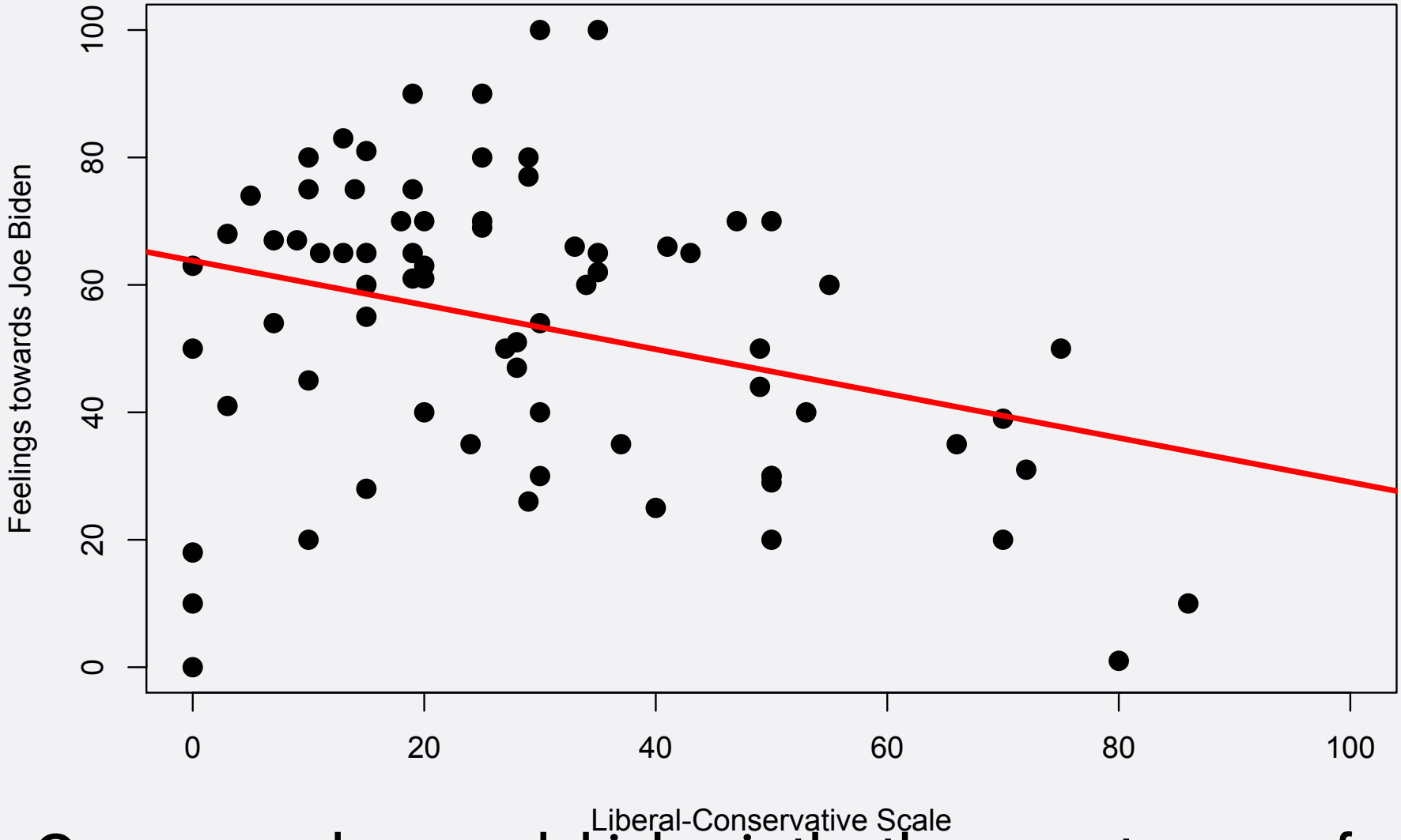
- r=-0.32
- Correlation: Direction and strength of relation, not size

JOE BIDEN



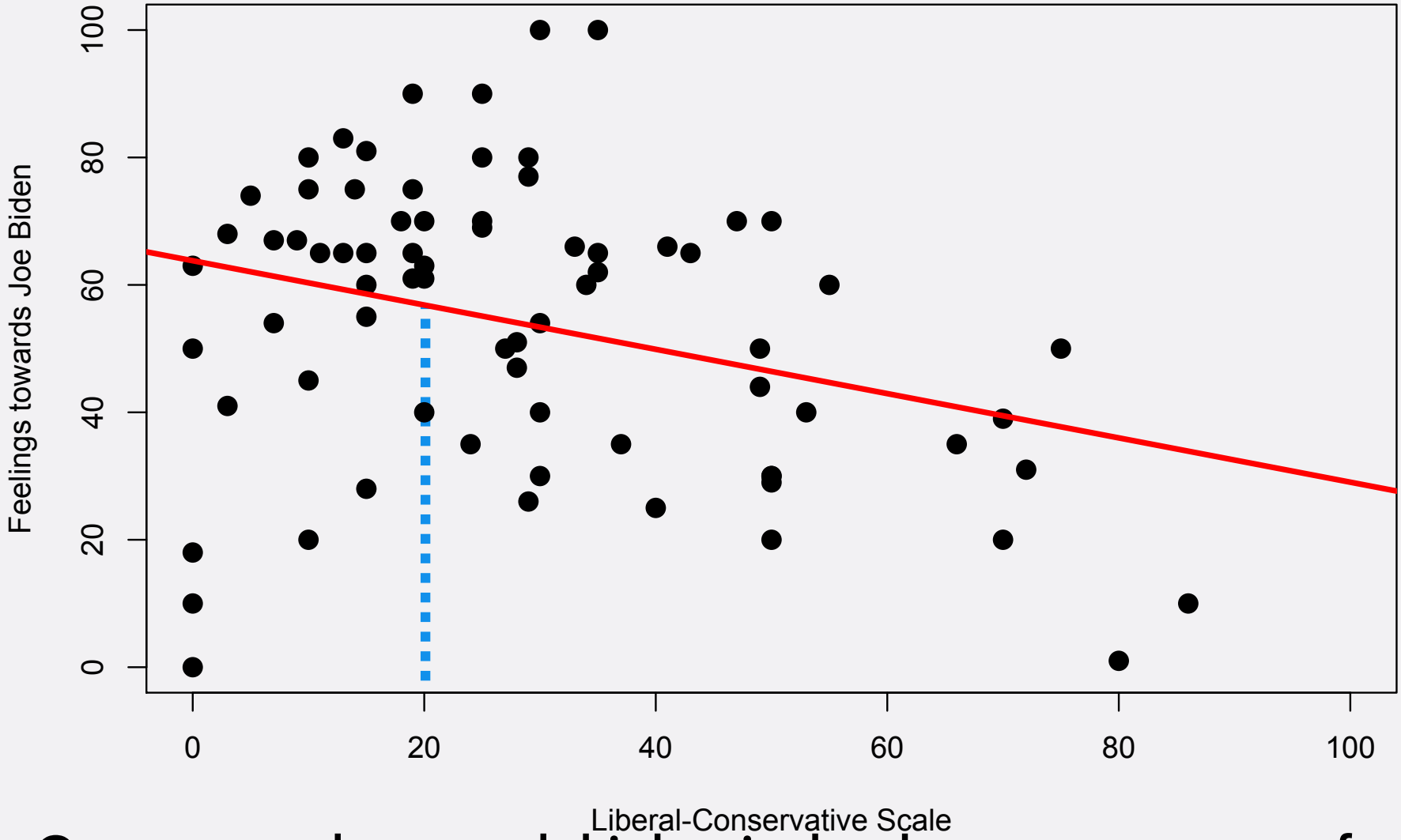
- **On average, how much higher is the thermometer score for someone who is a 20 on the liberal-conservative scale, compared to someone who is a 80?**

LINE



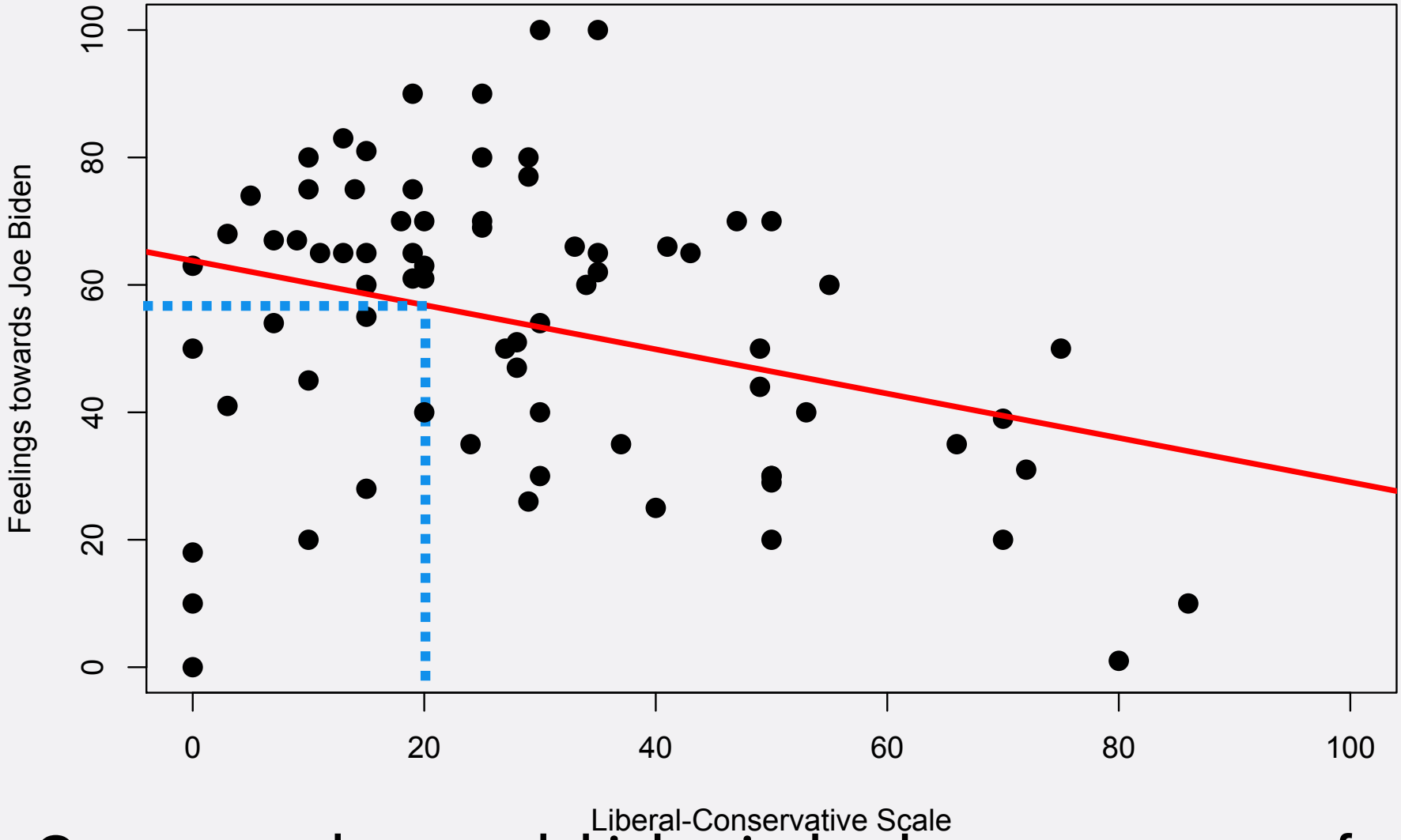
- **On average, how much higher is the thermometer score for someone who is a 20 on the liberal-conservative scale, compared to someone who is a 80?**

LINE



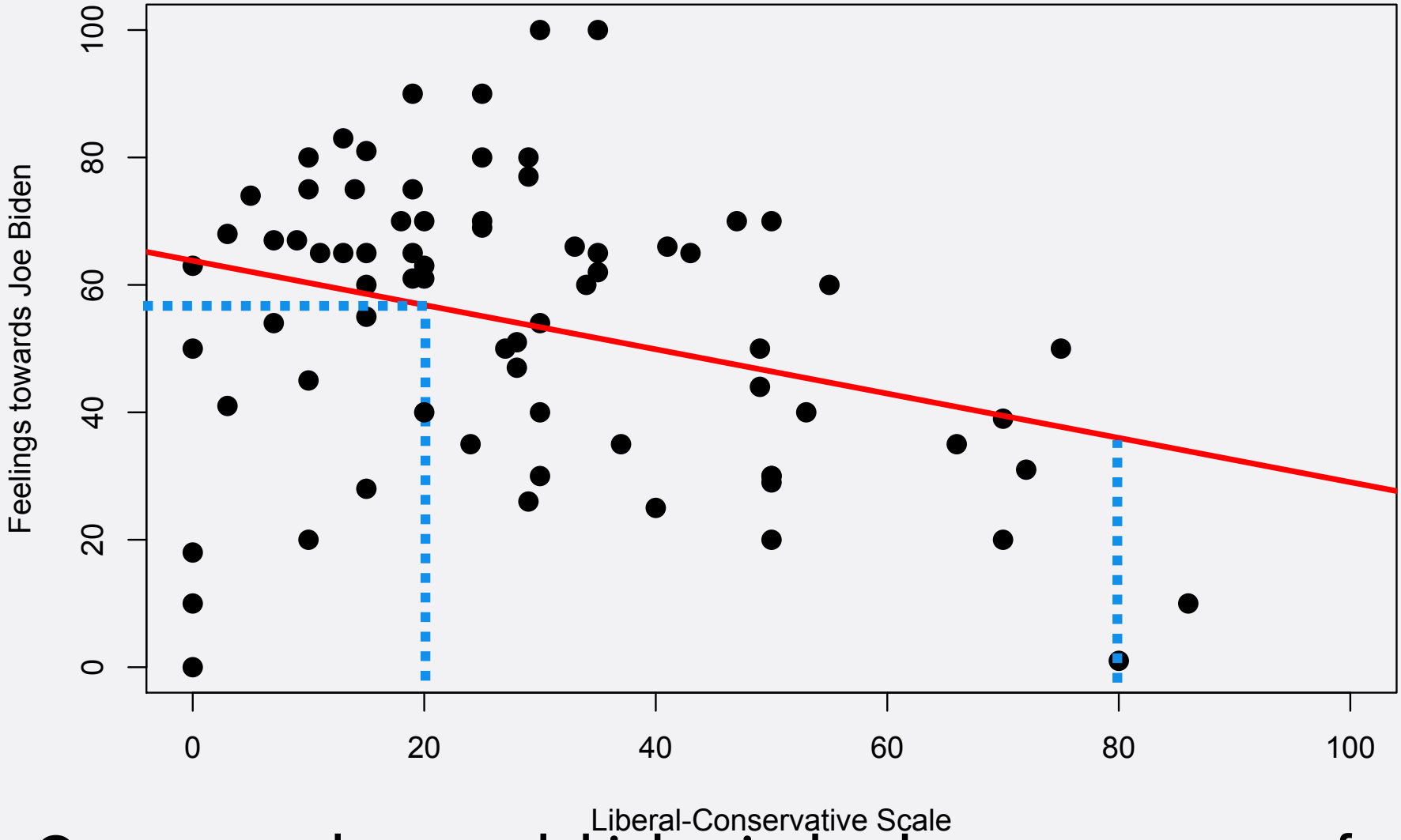
- **On average, how much higher is the thermometer score for someone who is a 20 on the liberal-conservative scale, compared to someone who is a 80?**

LINE



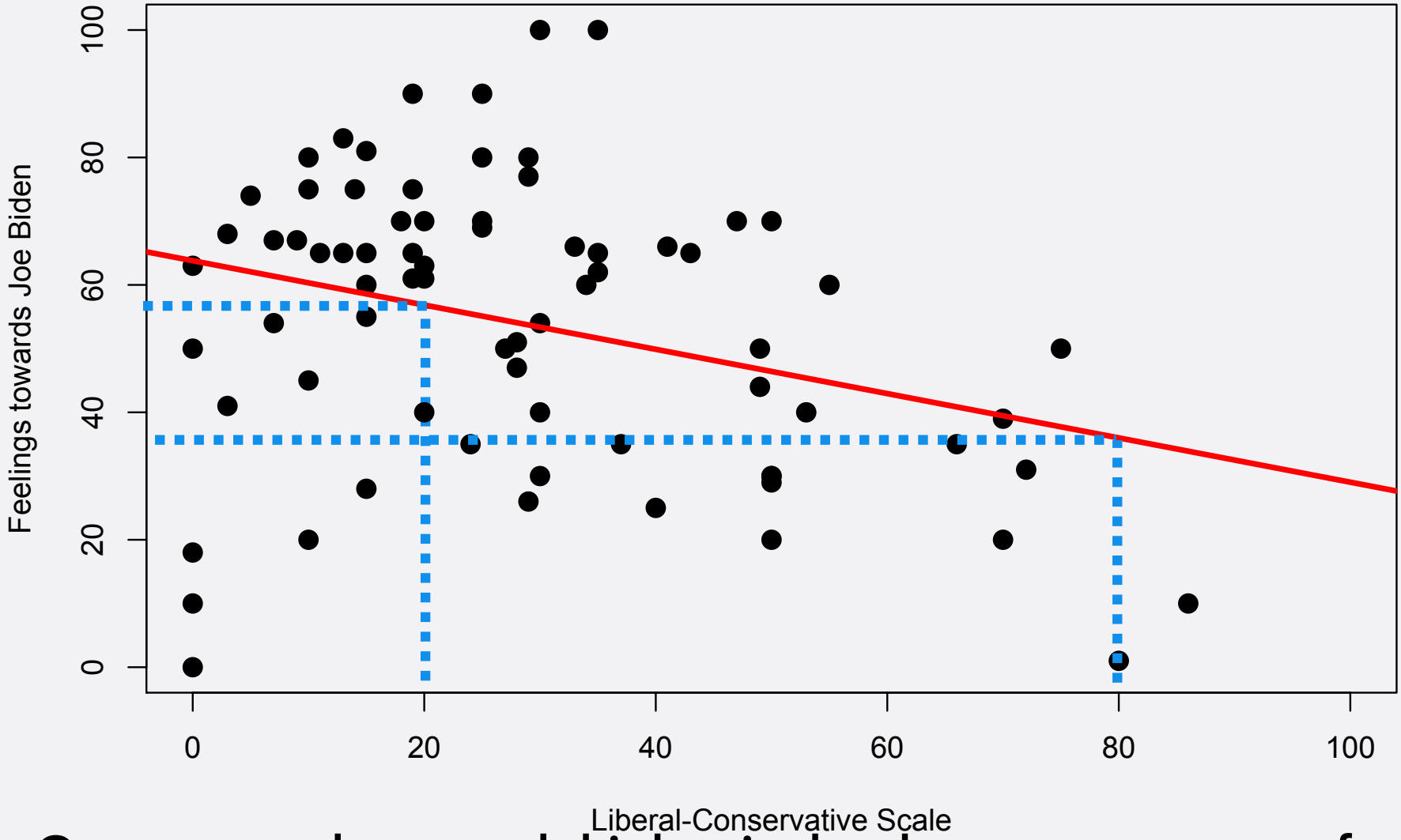
- **On average, how much higher is the thermometer score for someone who is a 20 on the liberal-conservative scale, compared to someone who is a 80?**

LINE



- **On average, how much higher is the thermometer score for someone who is a 20 on the liberal-conservative scale, compared to someone who is a 80?**

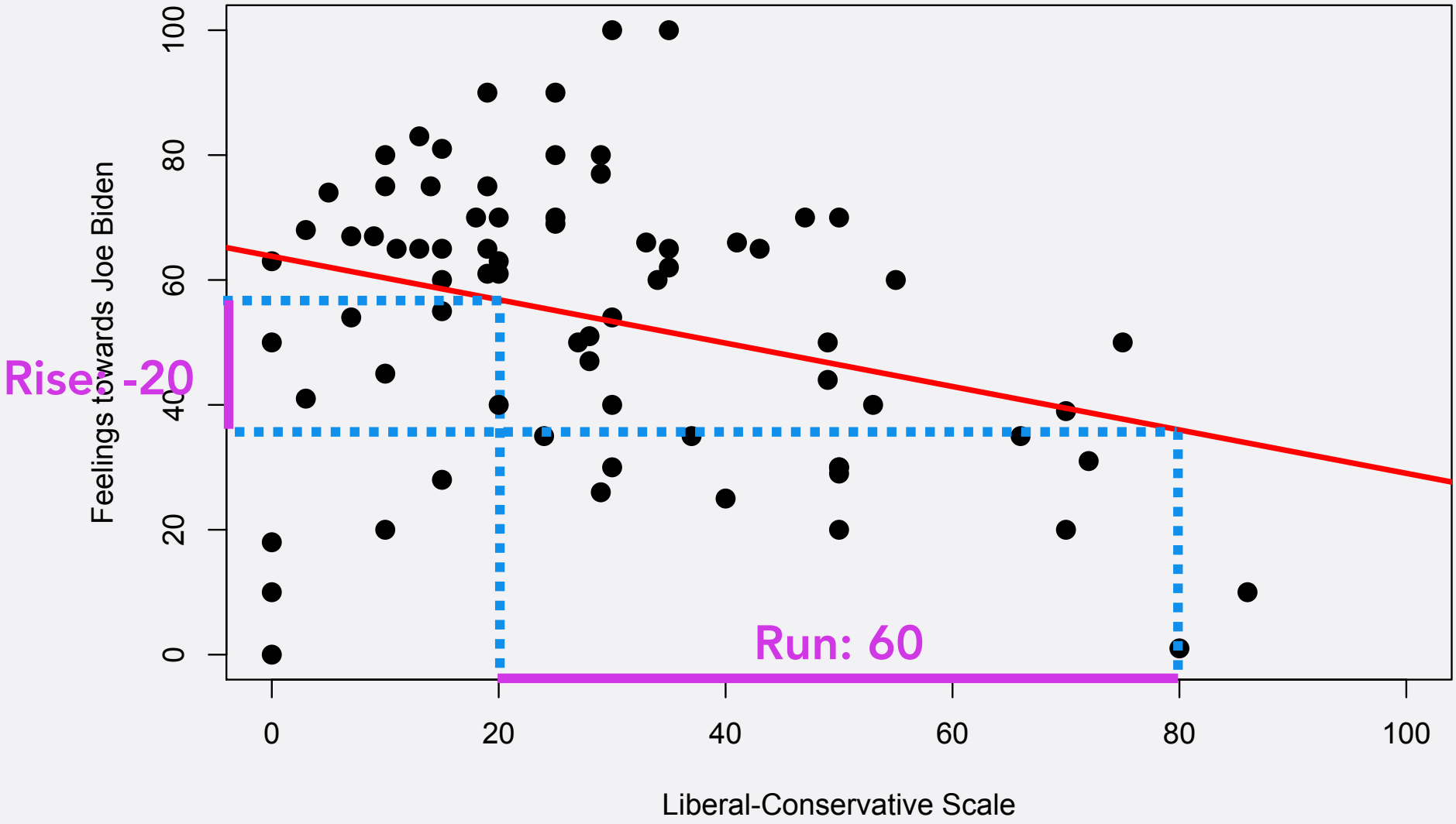
LINE



- **On average, how much higher is the thermometer score for someone who is a 20 on the liberal-conservative scale, compared to someone who is a 80?**

LINE

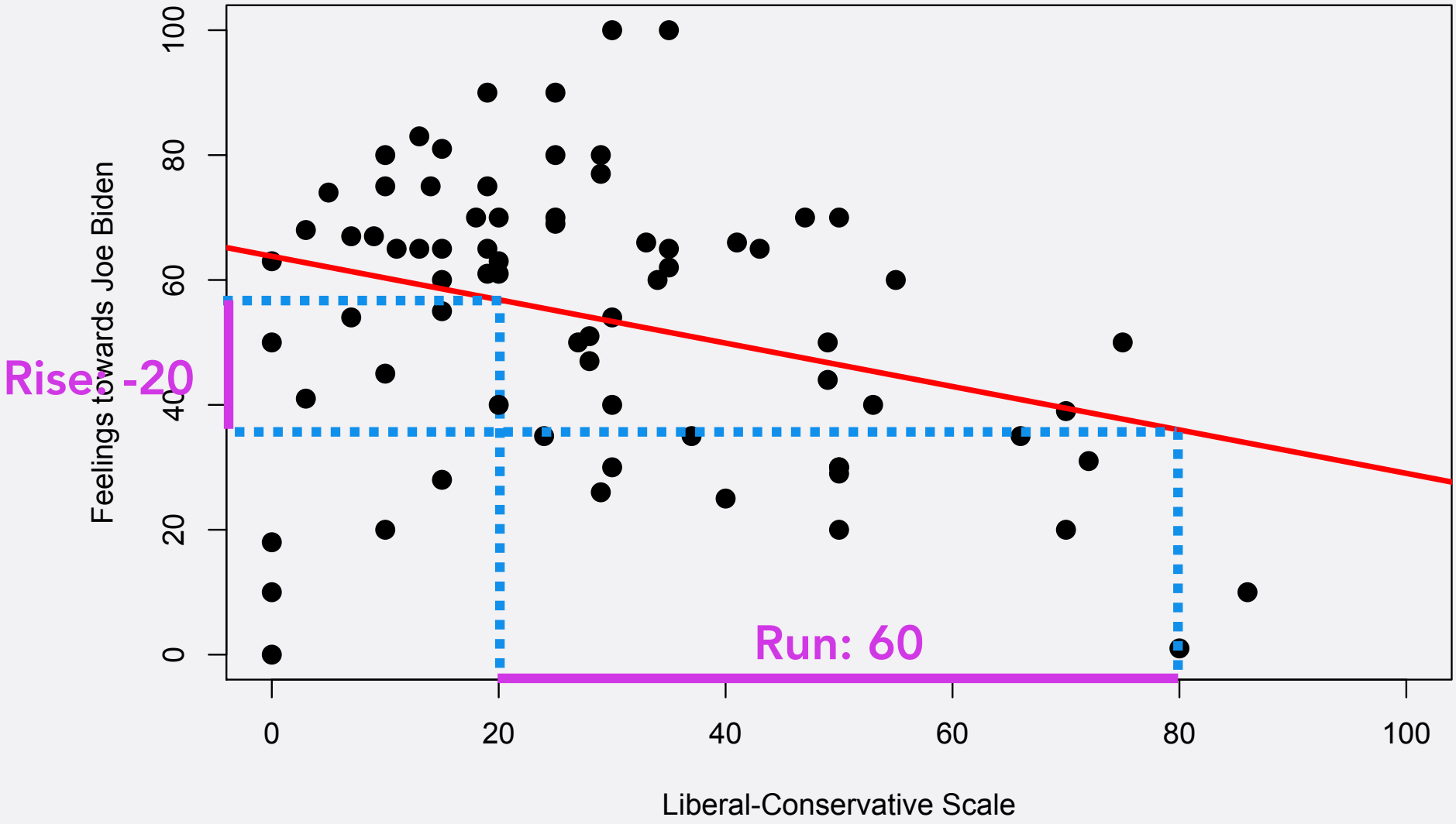
Slope=Rise over run



Slope=Rise over run=-20/60=-0.33

LINE

Slope=Rise over run



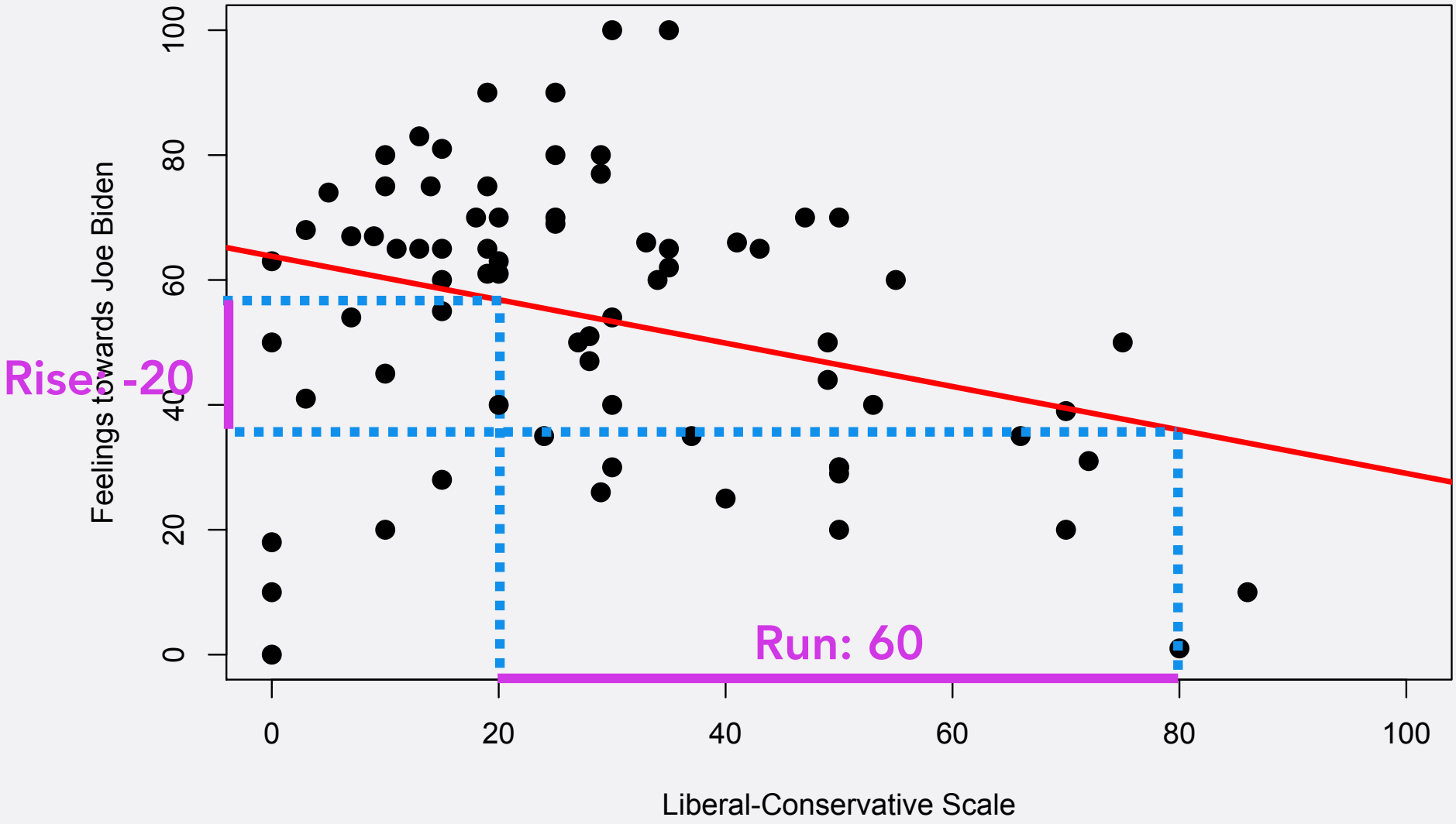
- For each one unit increase on the liberal-conservative scale, feelings towards J. Biden go down 0.33 points

NOTE

- In this case, it happens to be that slope is equal to correlation
- This does not need to be the case

LINE

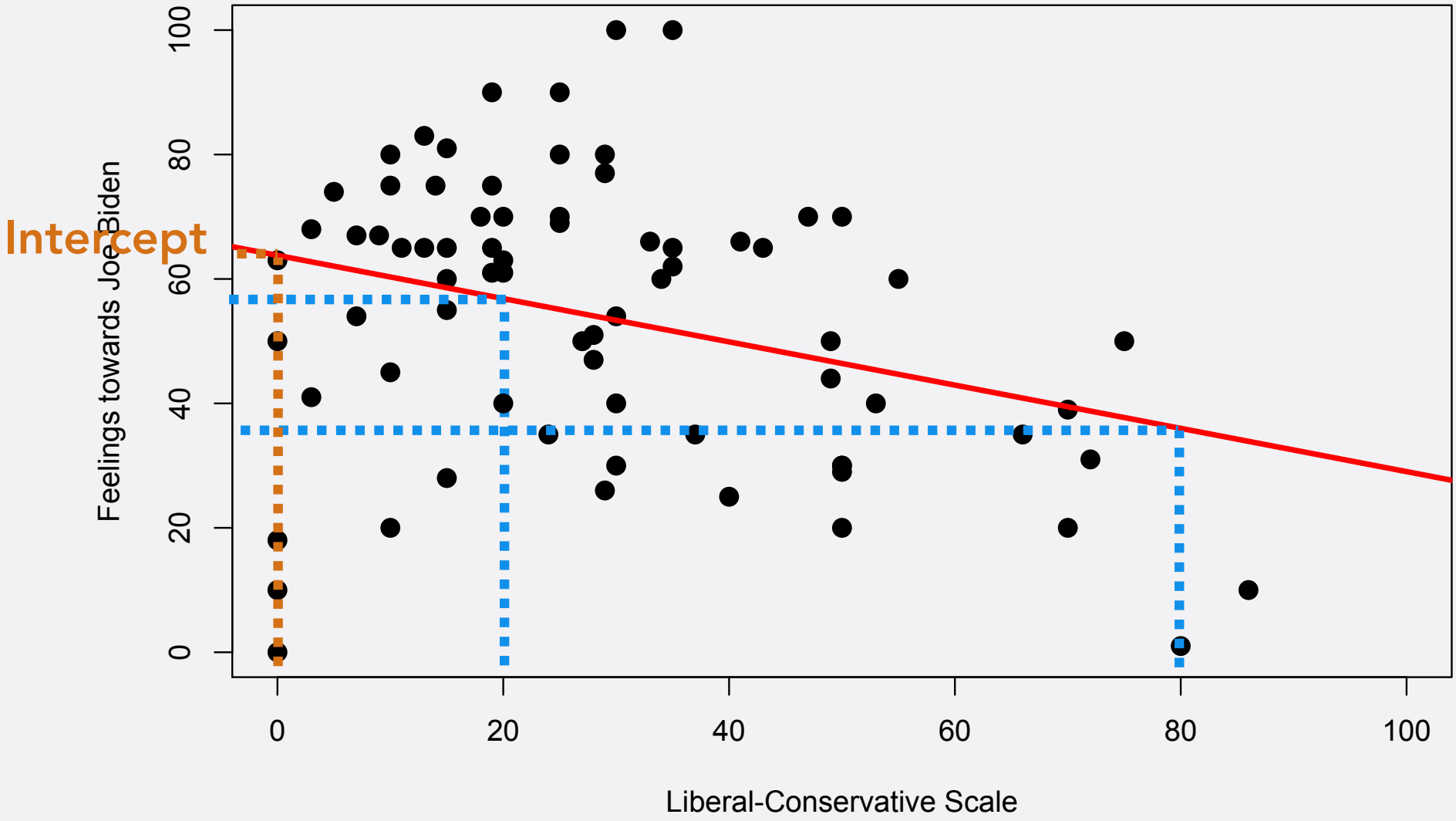
Slope=Rise over run



- For each one unit increase on the liberal-conservative scale, feelings towards J. Biden go down 0.33 points

LINE

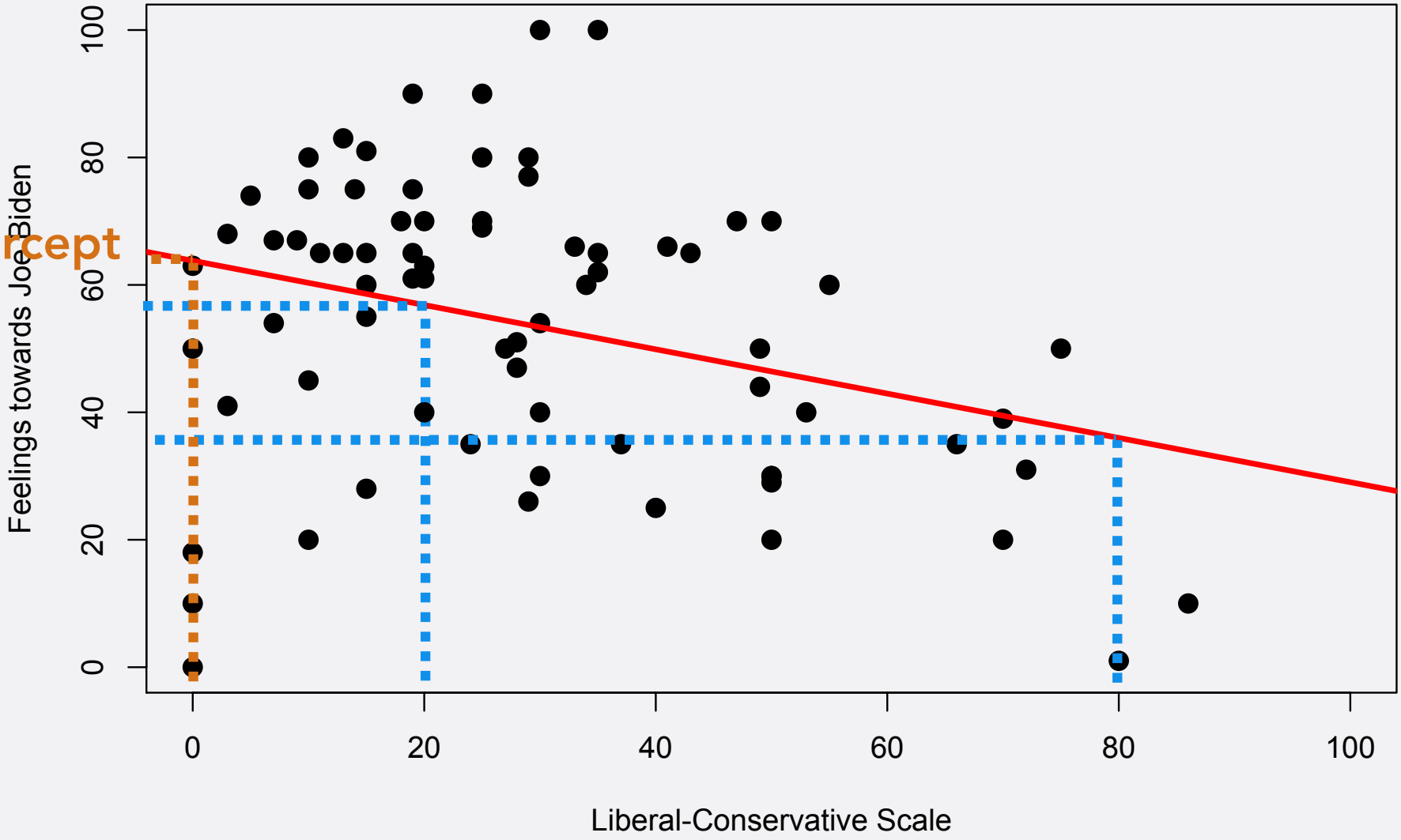
Slope=Rise over run



LINE

Slope=Rise over run

Intercept



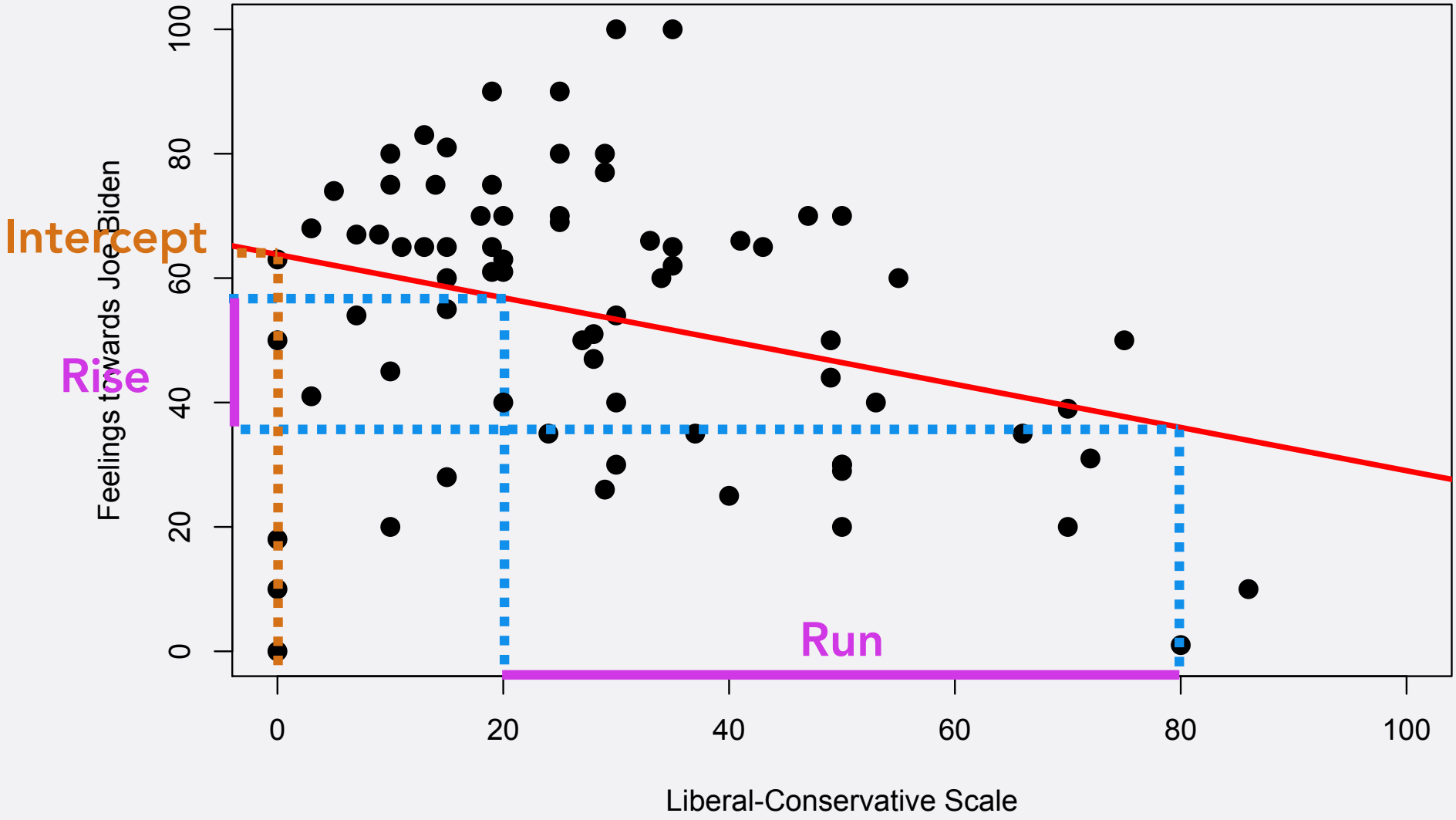
- **Students who are very liberal (score=0) are expected to have a feeling thermometer score of 64.**

LINEAR REGRESSION

- Linear regression: Equation that tells us *direction* and *size* of relationship between independent variable (IV) and dependent variable (DV)
- DV = Intercept + Slope * IV

LINE

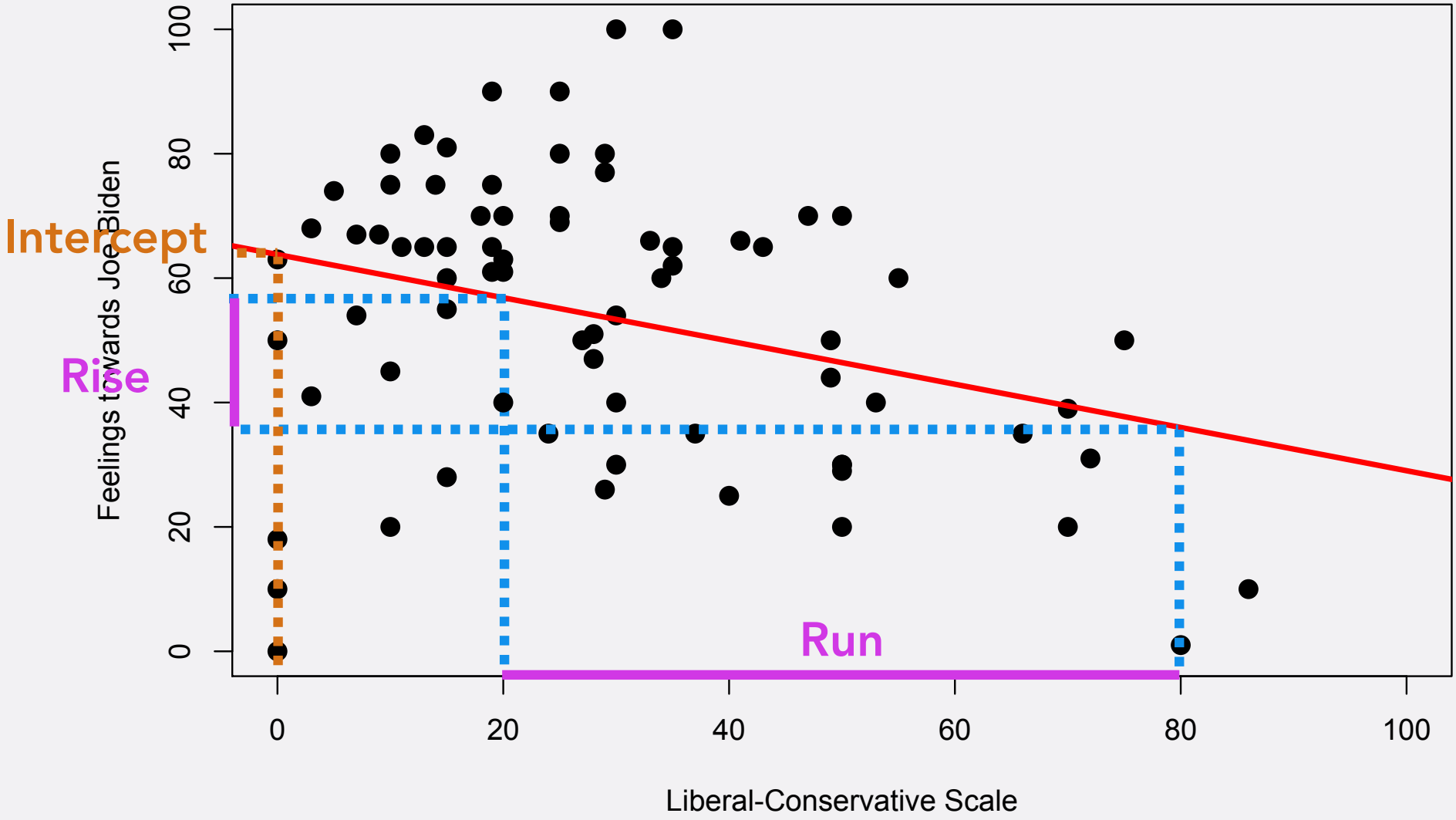
Slope=Rise over run



- Thermometer Score = Intercept + Slope * Lib/Cons

LINE

Slope=Rise over run



- Thermometer Score = 64 - 0.33 * Lib/Cons

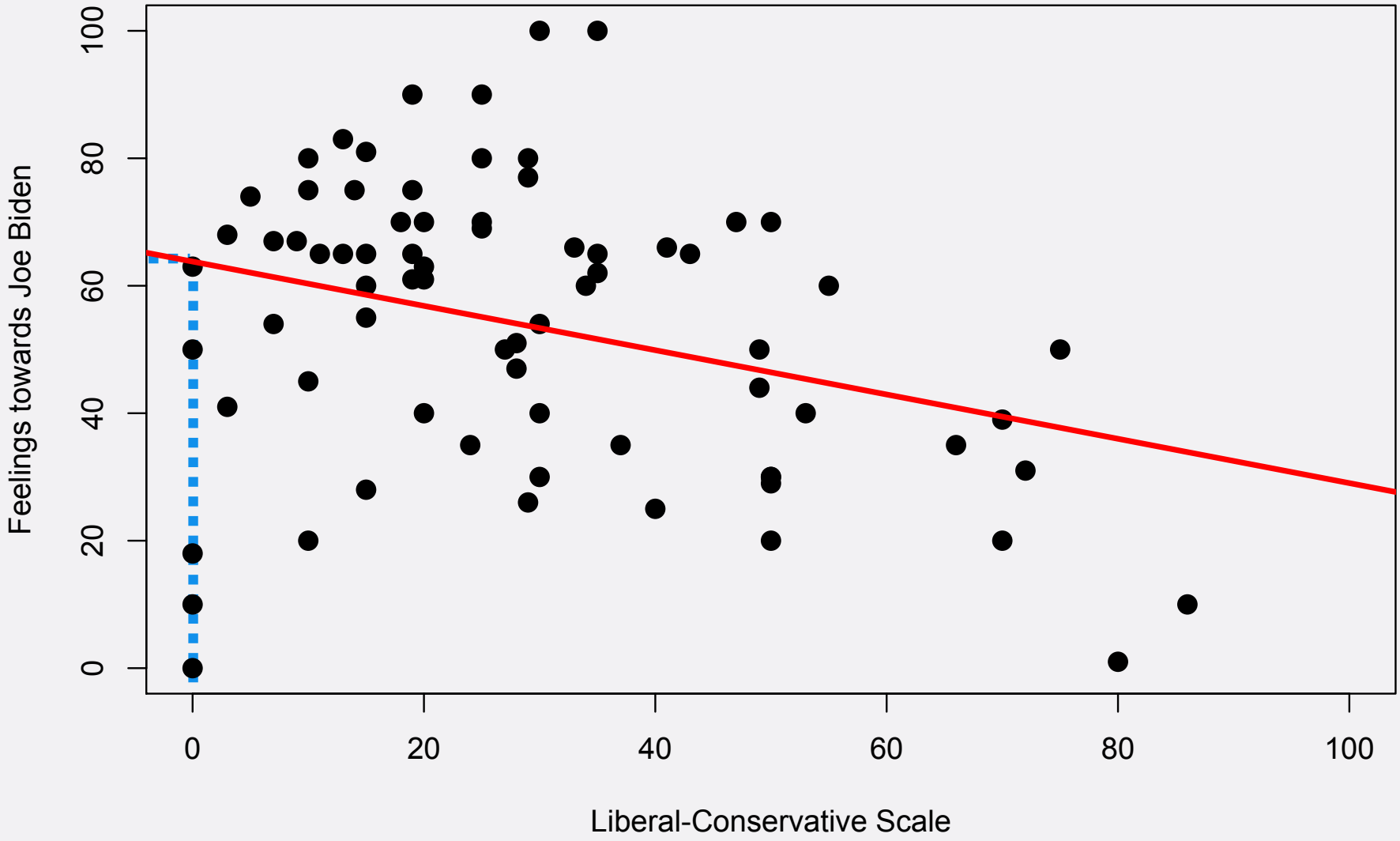
WHAT THIS TELLS US

- Thermometer Score = 64 - 0.33 * Lib/Cons
- Can predict what someone's thermometer rating of Joe Biden will be, depending on where they are on liberal-conservative scale

WHAT THIS TELLS US

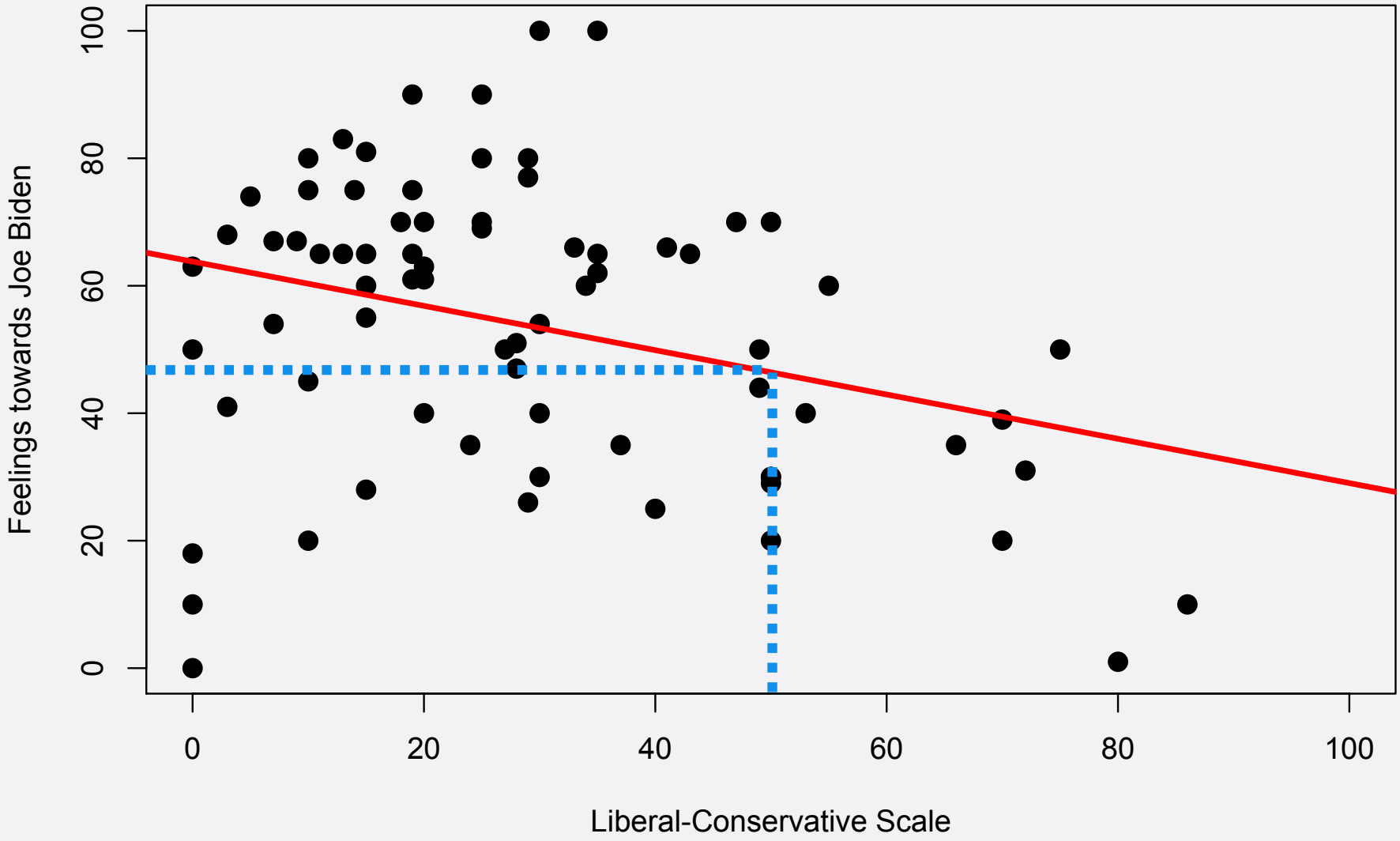
- Thermometer Score = 64 - 0.33 * Lib/Cons
- Lib/Cons scale of 0:
  - 64 - 0.33 * 0 = 64

LINE



- **64 - 0.33 * 0 = 64**

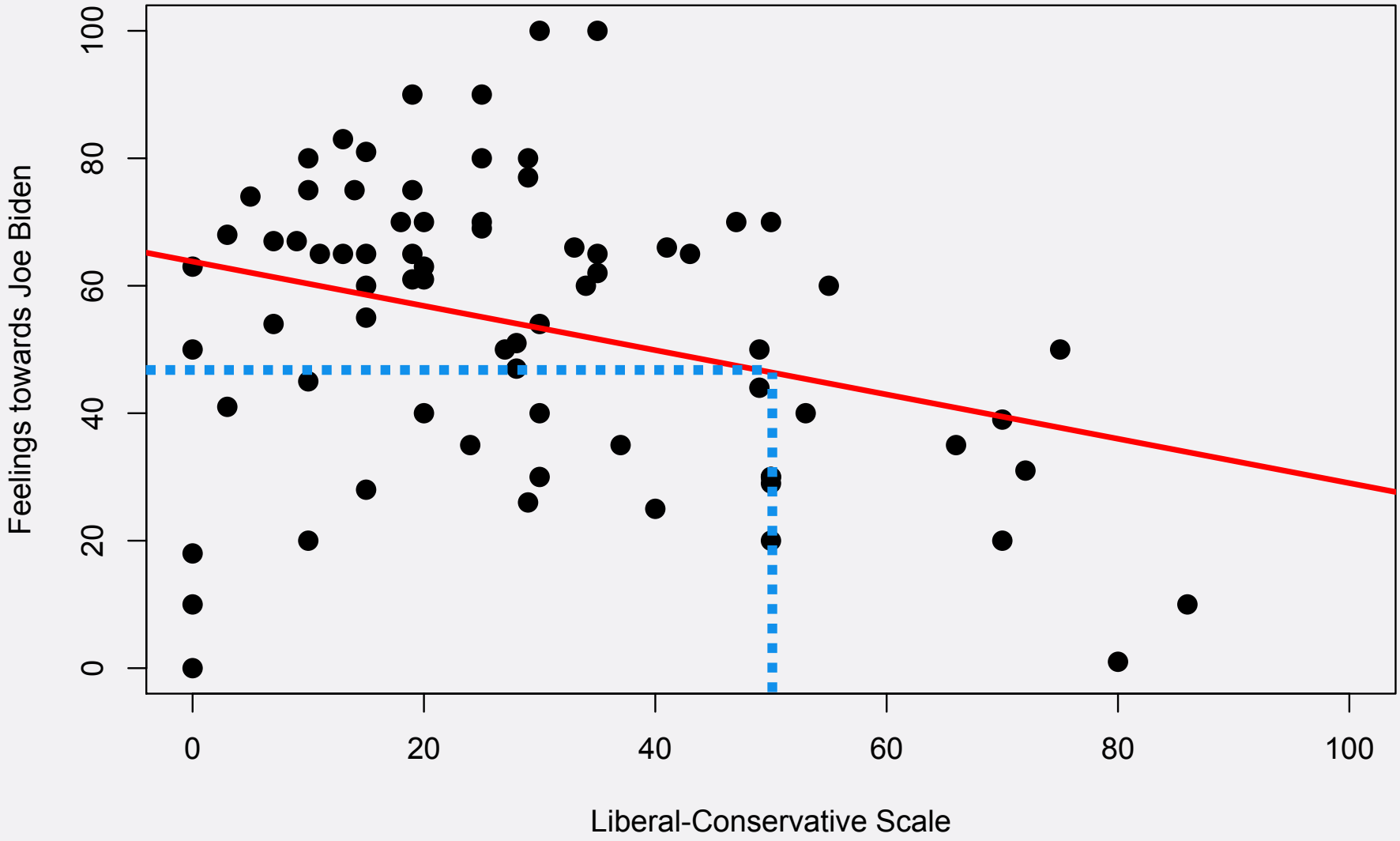
LINE


Feelings towards Joe Biden
0
20
40
60
80
100
Liberal-Conservative Scale
0
20
40
60
80
100

WHAT THIS TELLS US

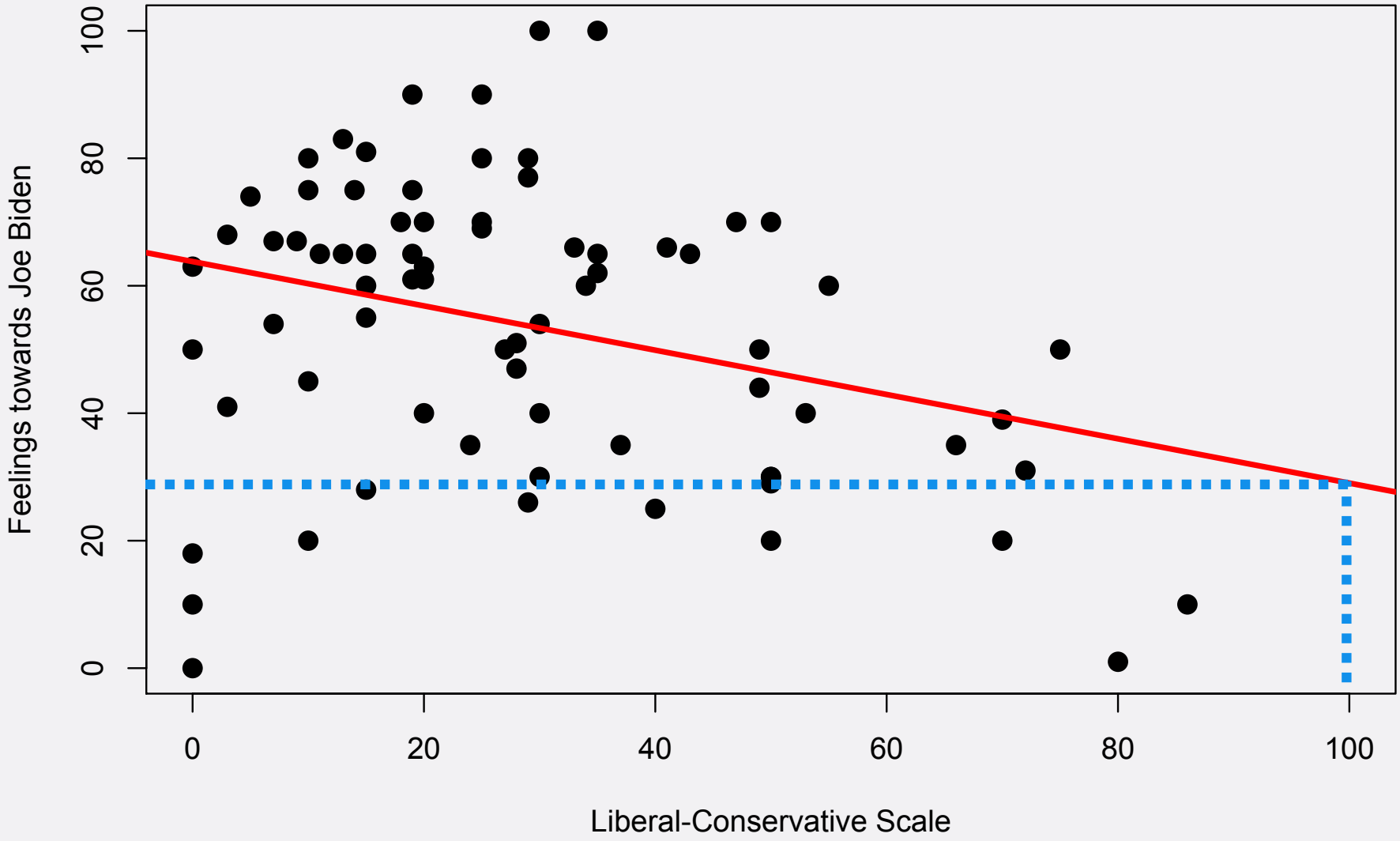
- Thermometer Score = 64 - 0.33 * Lib/Cons
- Lib/Cons scale of 50:
  - 64 - 0.33 * 50 = 47.5

LINE



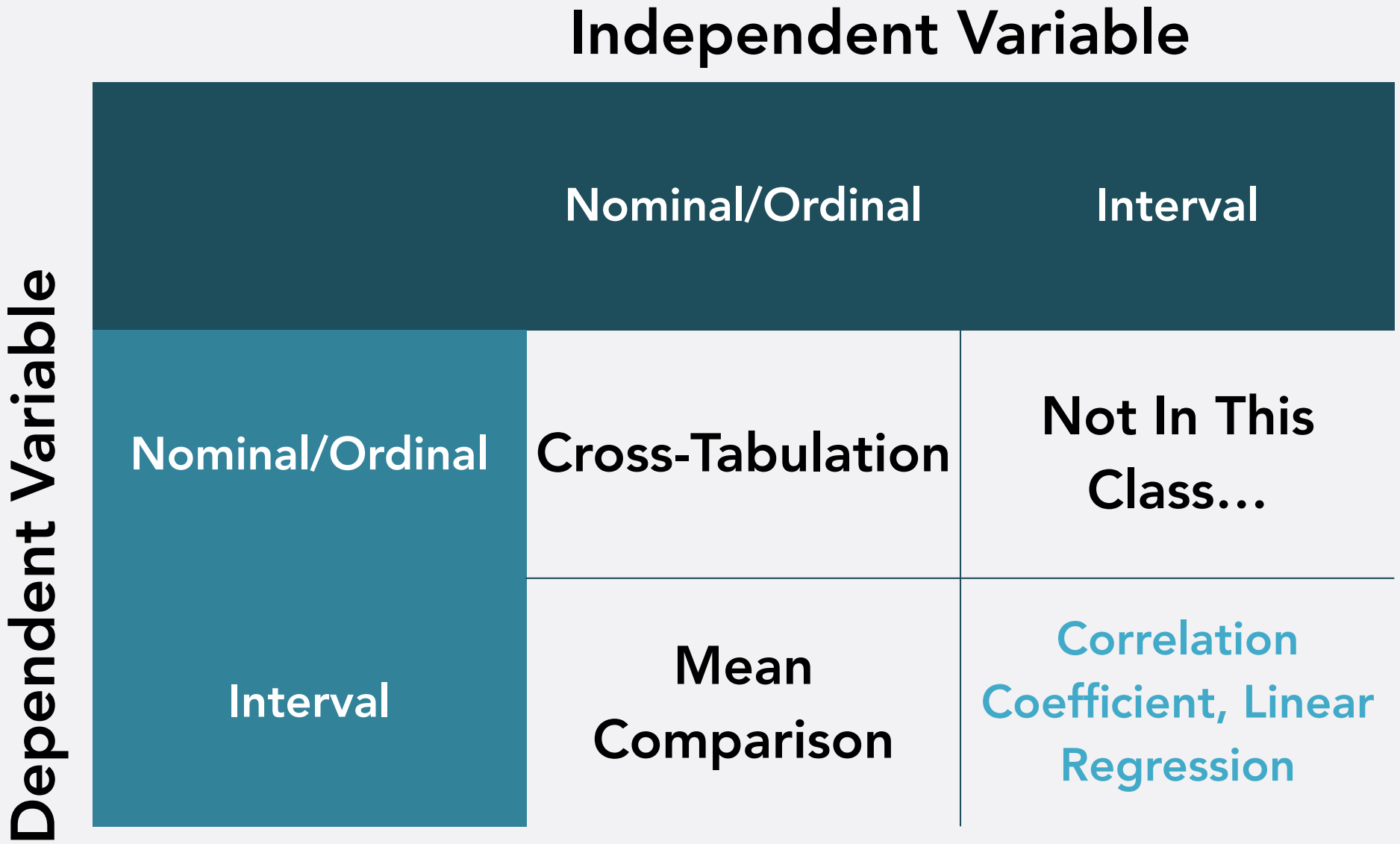
- **64 - 0.33 * 50 = 47.5**

LINE



- **64 - 0.33 * 100 = 31**

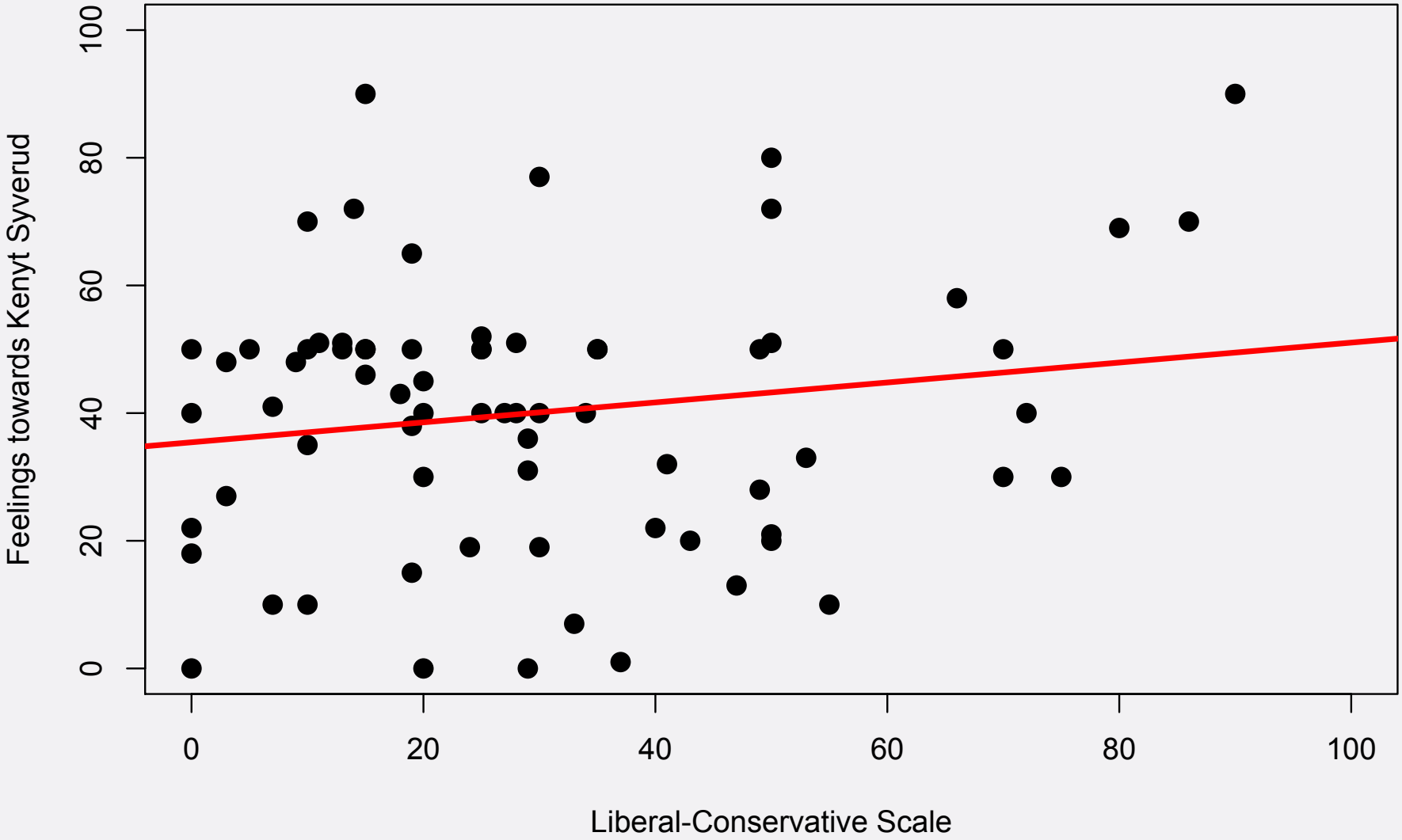
BIVARIATE RELATIONSHIPS



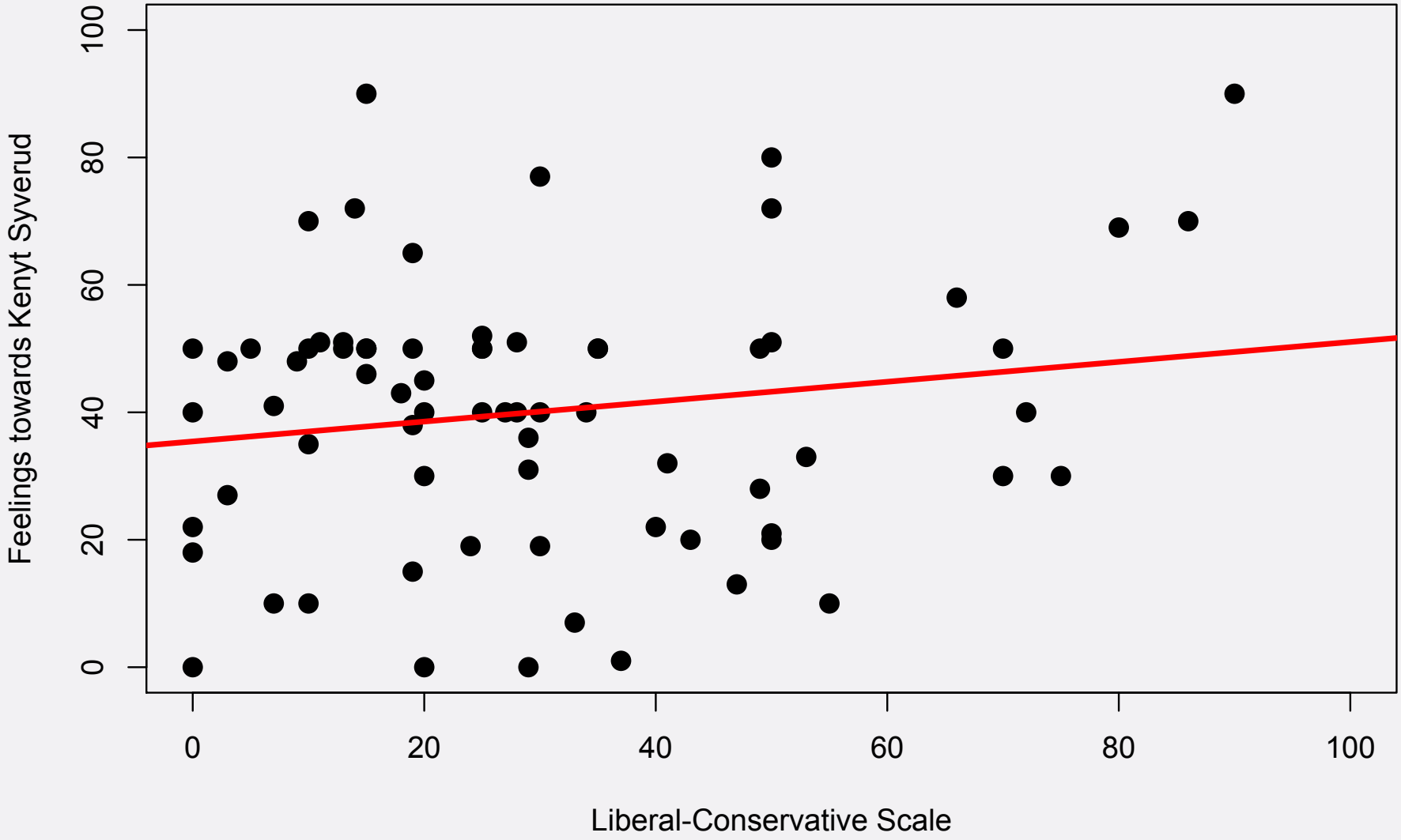
LINEAR REGRESSION

- A tool that tells us the direction and *size* of the effect of an independent variable on a dependent variable
  - both are interval-level

KENT SYVERUD

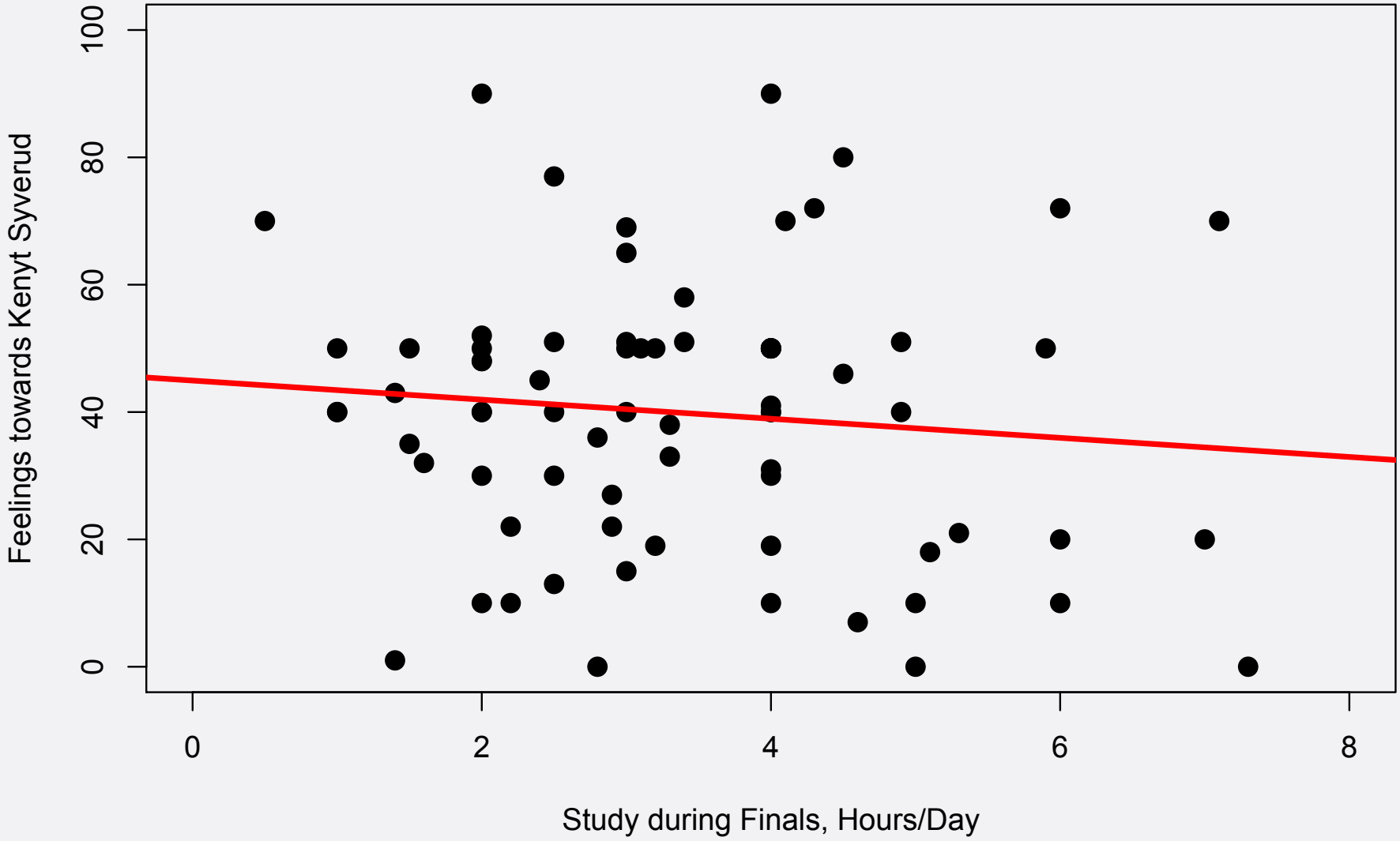


KENT SYVERUD



- Thermometer Score = 35 + 0.16 * Lib/Cons

DIFFERENT INDEPENDENT VARIABLE



- Thermometer Score = 55 - 2.6 * Hours/Day

INTERPRETATION?

- Thermometer Score = 55 - 2.6 * Hours/Day
  - What does the 55 tell us?
  - What does the -2.6 tell us?

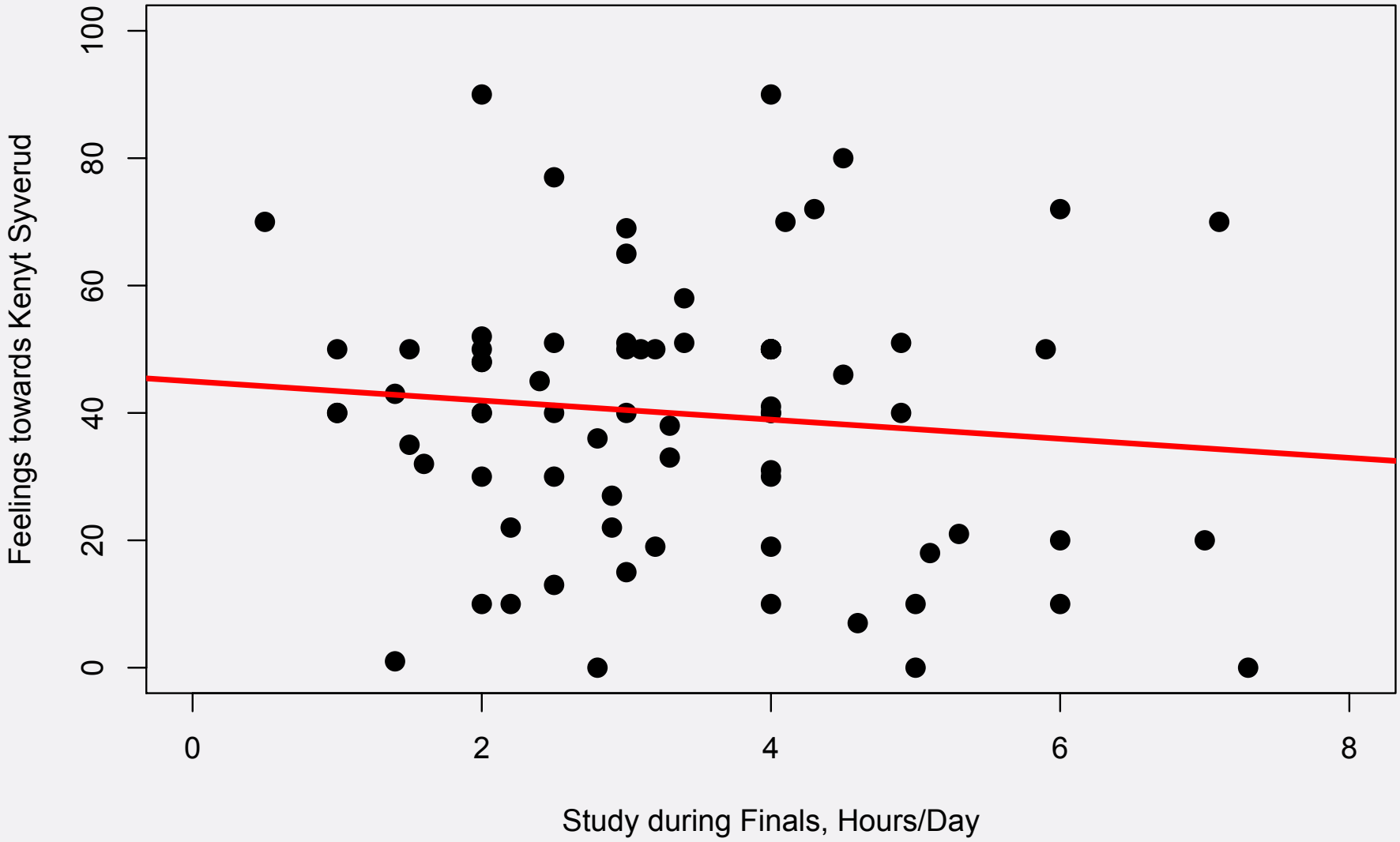
INTERPRETATION?

- Thermometer Score = 55 - 2.6 * Hours/Day
  - What does the 55 tell us?
    - A student who studies 0 hours per day has an expected thermometer score of 55
  - What does the -2.6 tell us?
    - For every one hour a student studies longer per day, their thermometer score is expected to decrease by 2.6 points

INTERPRETATION?

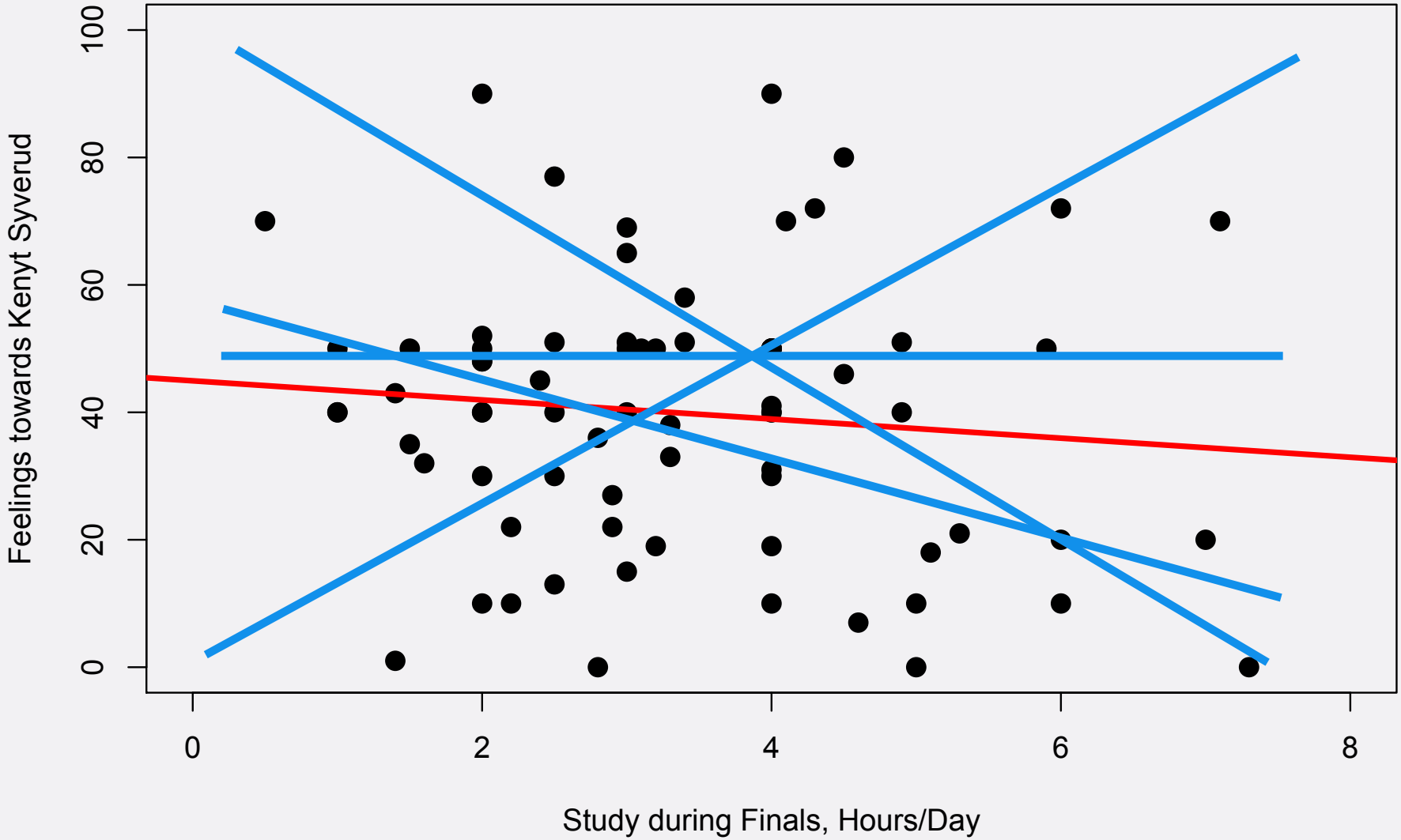
- Earlier we had:
  - Thermometer Score = 35 + 0.16 * Lib/Cons
- Now we have:
  - Thermometer Score = 55 - 2.6 * Hours/Day
- Does this mean that the effect of hours of study is larger than of how liberal-conservative students are?

HOW TO PICK THE LINE



- **Why this line?**

HOW TO PICK THE LINE

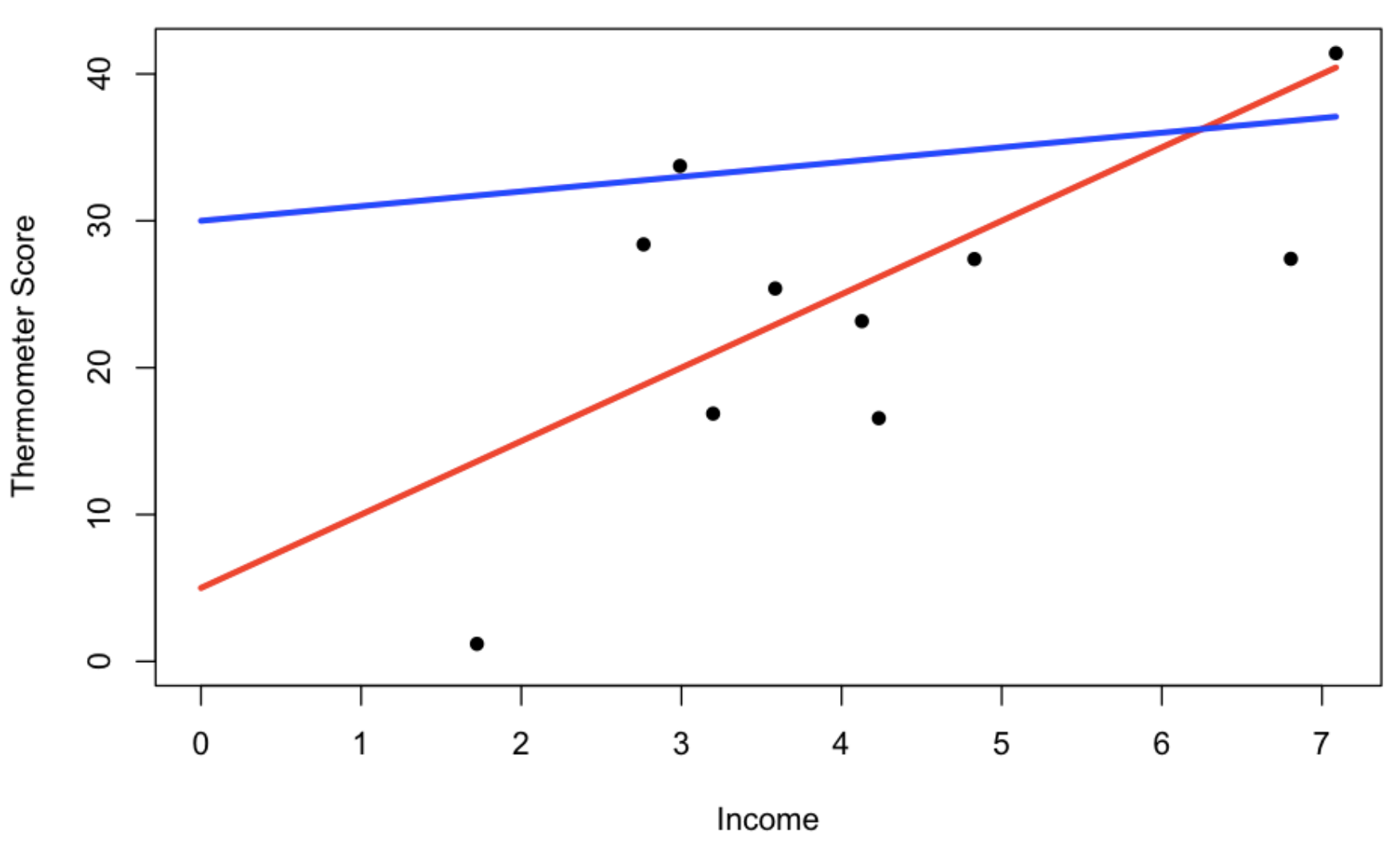


- **Why not these?**

MORE ON REGRESSION LINE

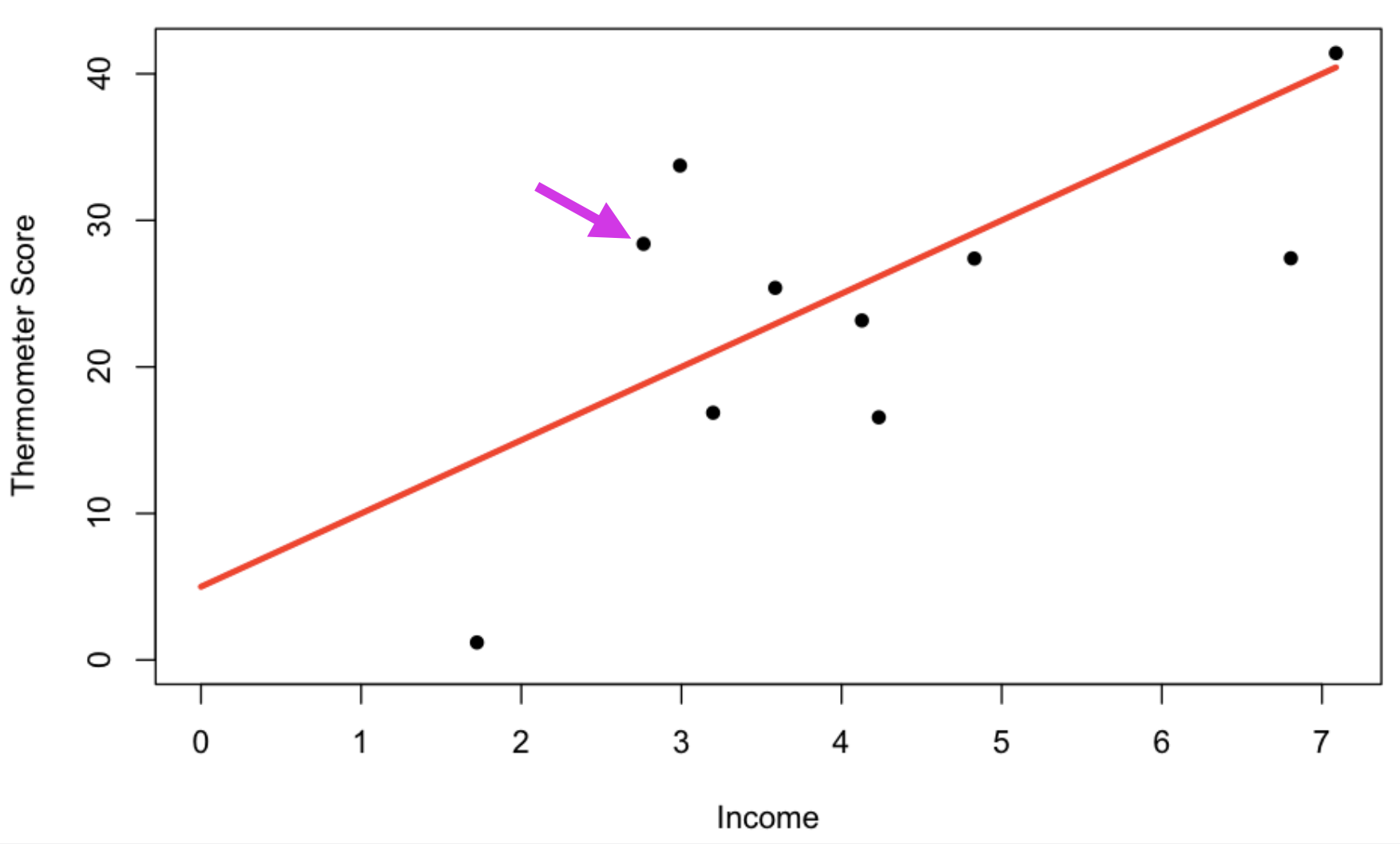
- How do I pick the line?
- How is linear regression useful?
- Caveats about linear regression

HOW TO PICK THE LINE

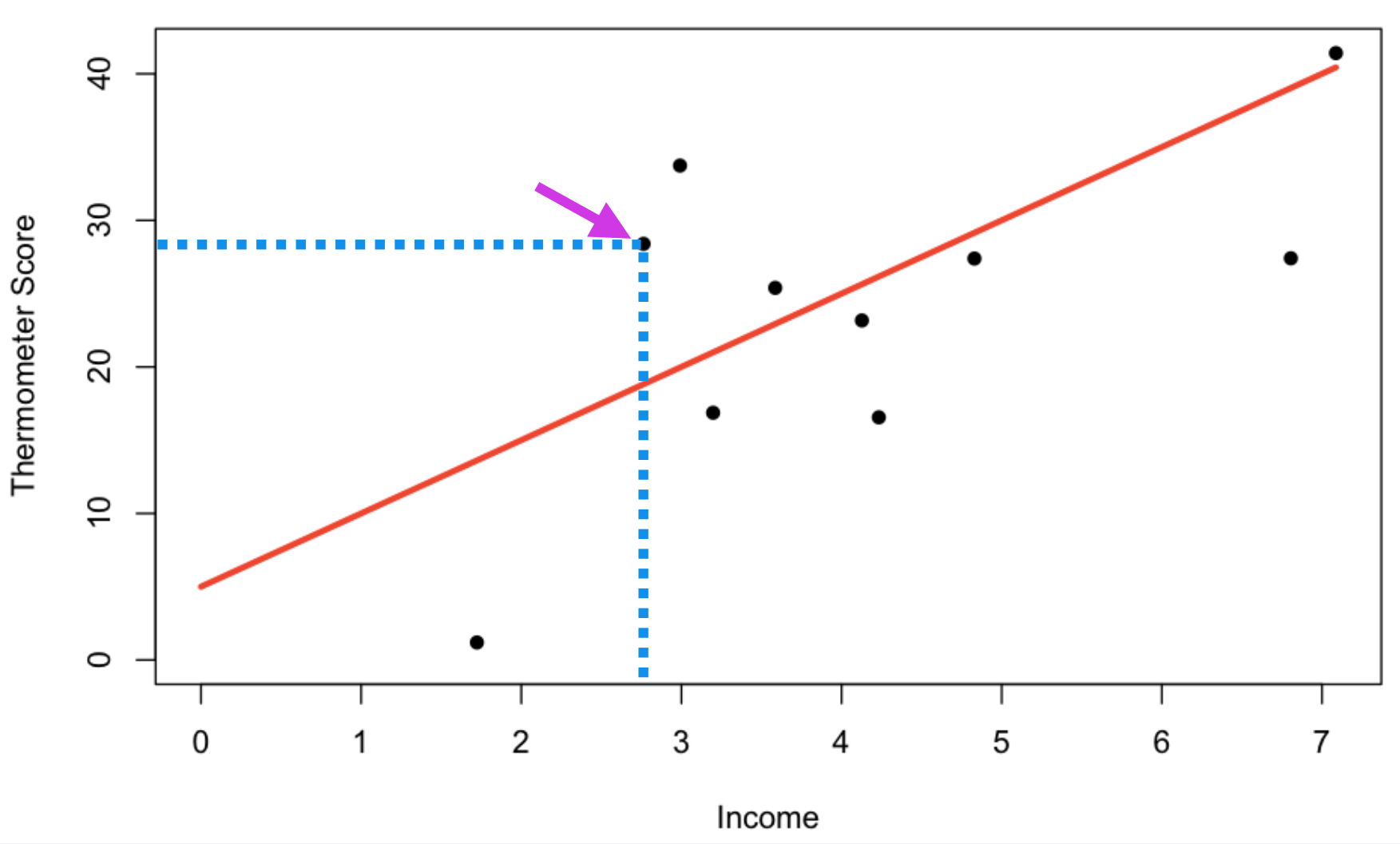


- **Which line is better?**

HOW TO PICK THE LINE

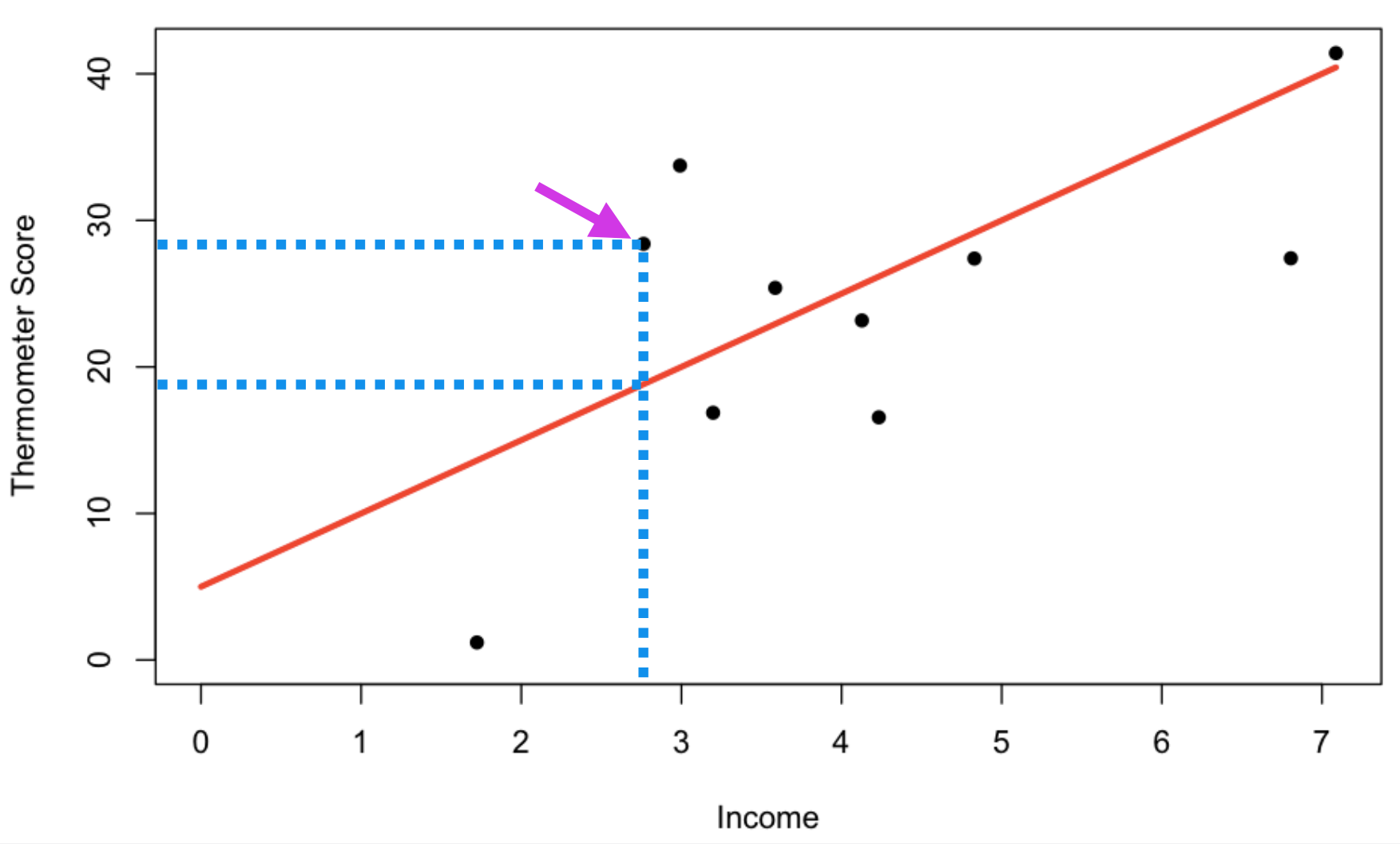


HOW TO PICK THE LINE



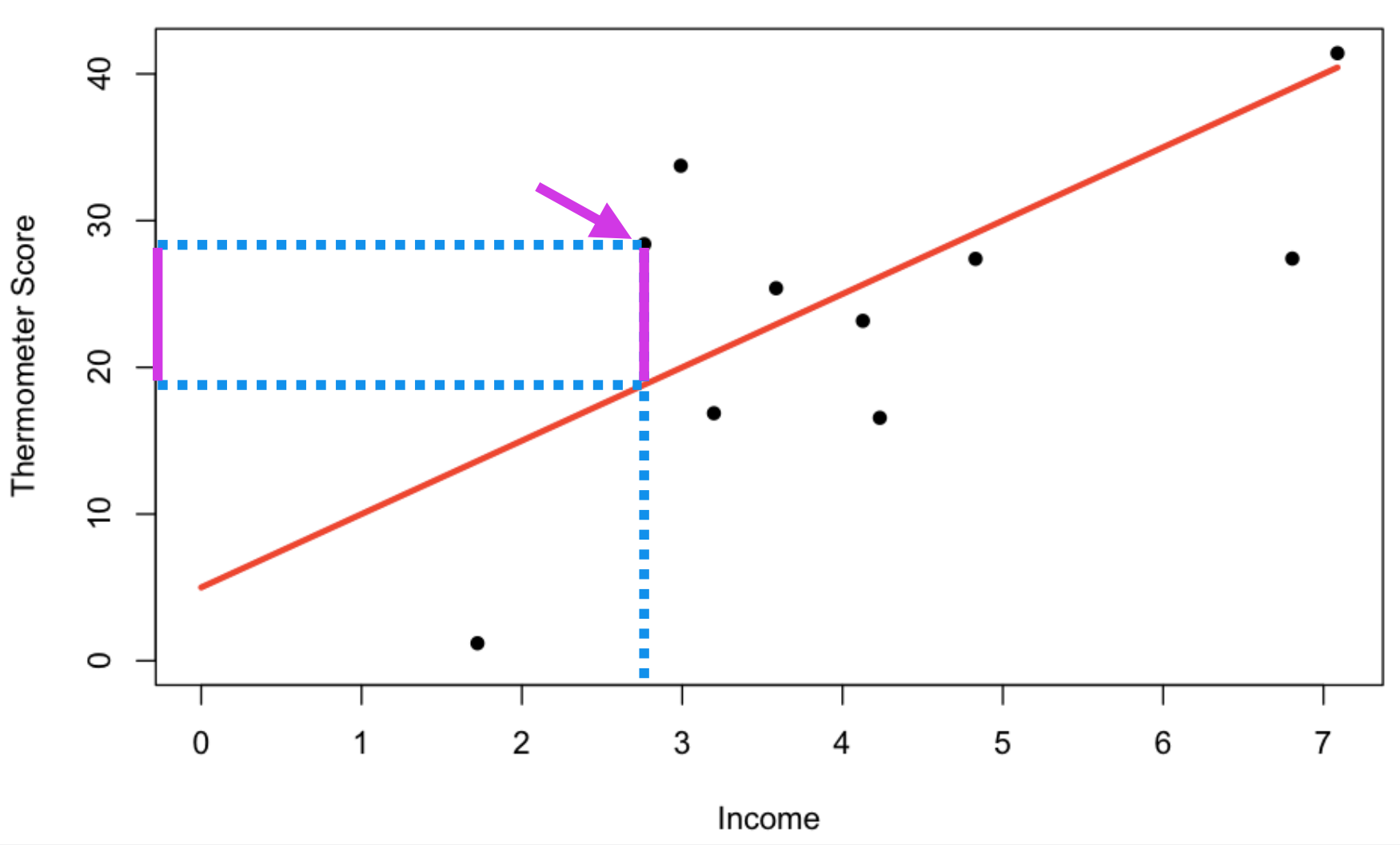
- **Actual y-value: y=28**

HOW TO PICK THE LINE



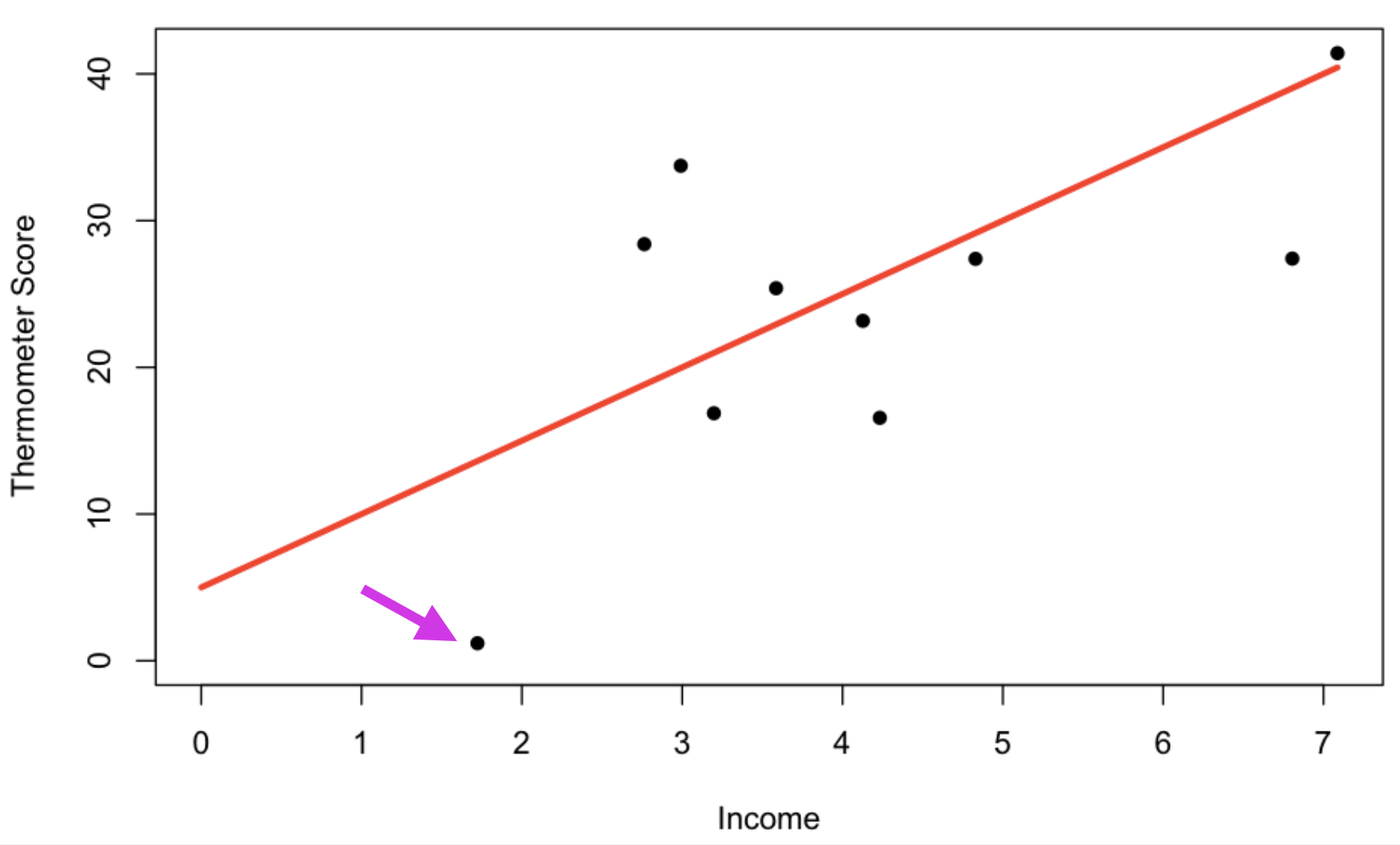
- **Predicted y-value: $\hat{y}$=19**

HOW TO PICK THE LINE

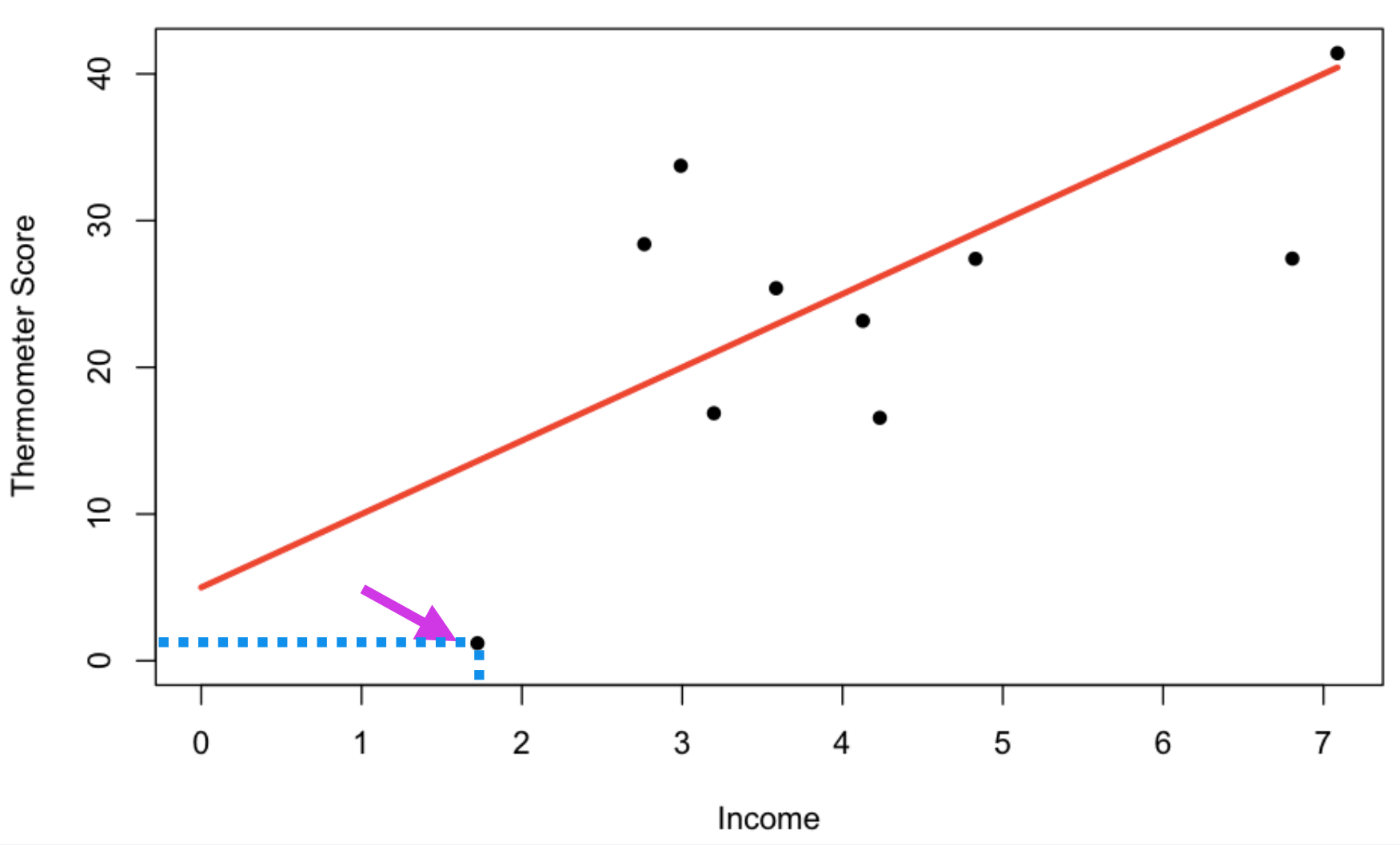


- **Prediction error: y - ŷ = 28 - 19 = 9**

HOW TO PICK THE LINE

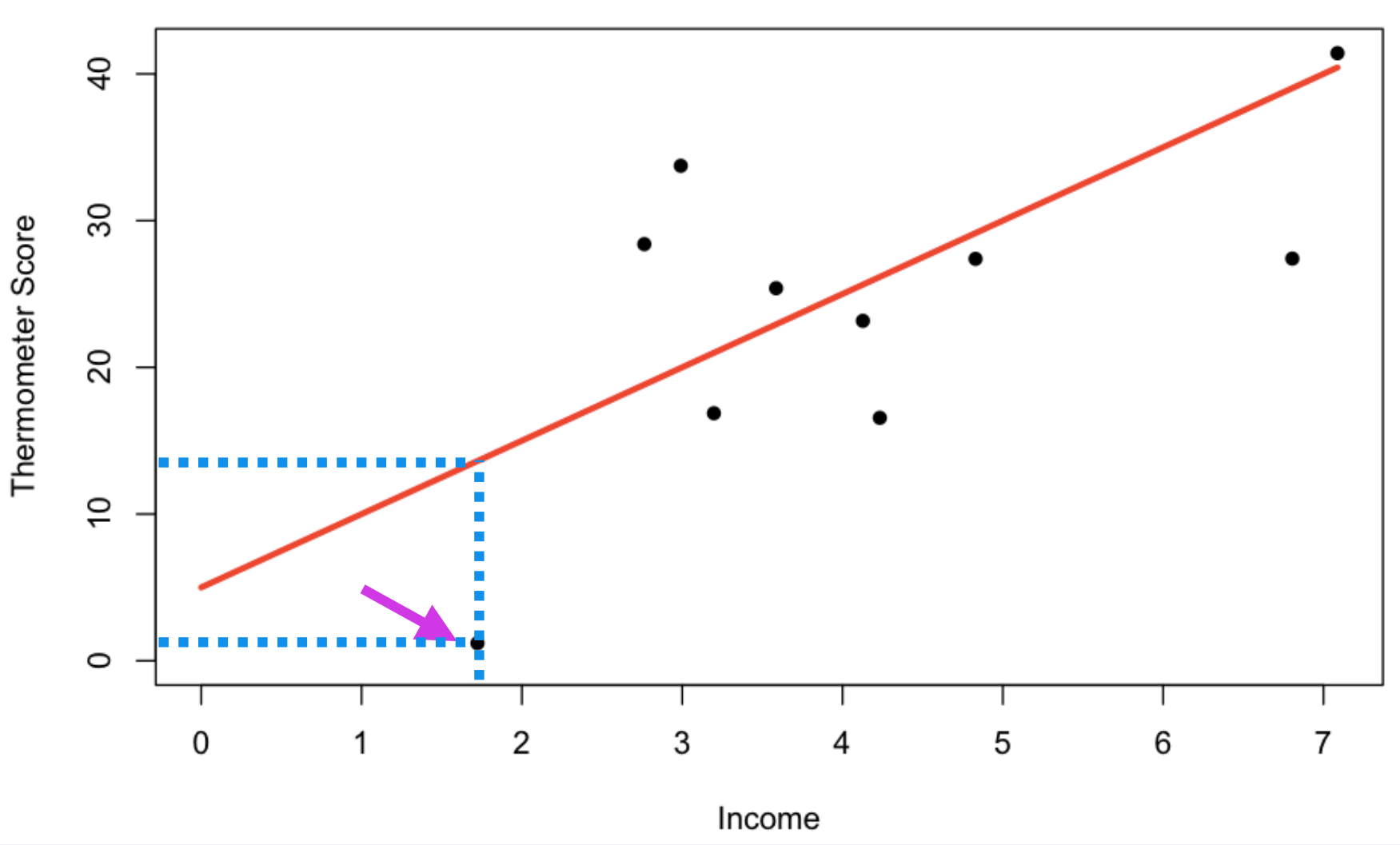


HOW TO PICK THE LINE



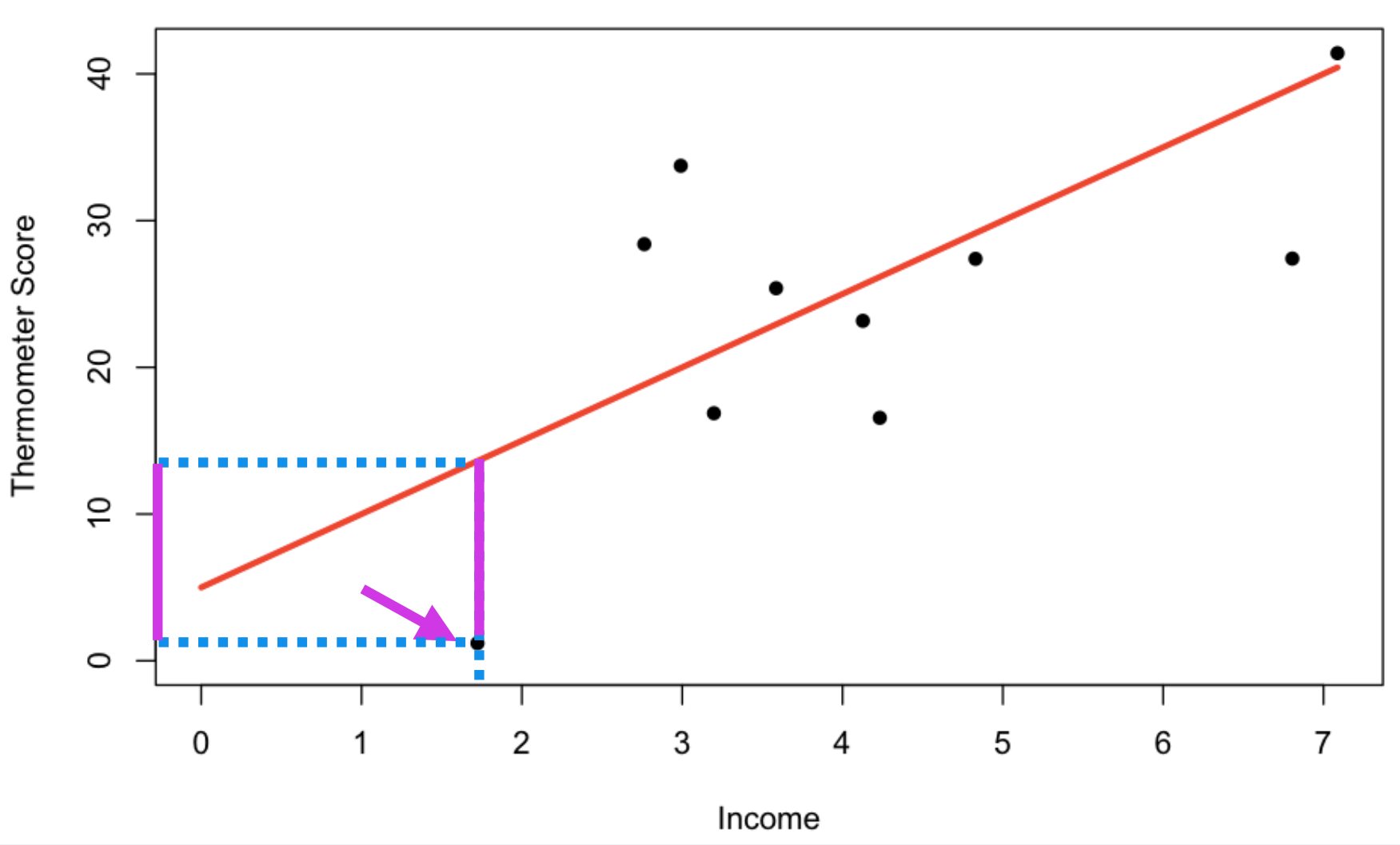
- **Actual y-value: y=1**

HOW TO PICK THE LINE



- **Predicted y-value: $\hat{y}=14$**

HOW TO PICK THE LINE

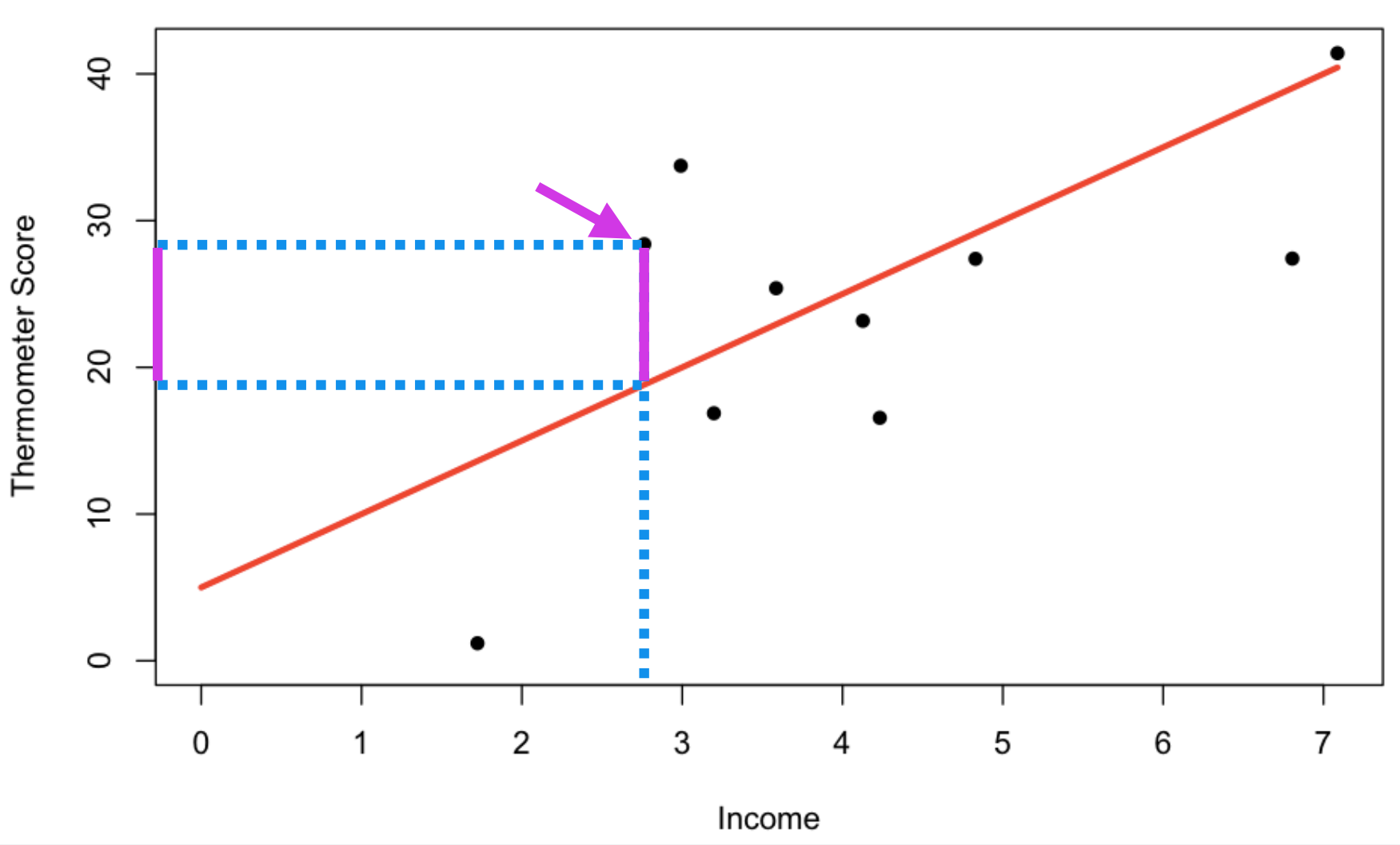


- **Prediction error: $y - \hat{y} = 1 - 14 = -13$**

PREDICTION ERROR

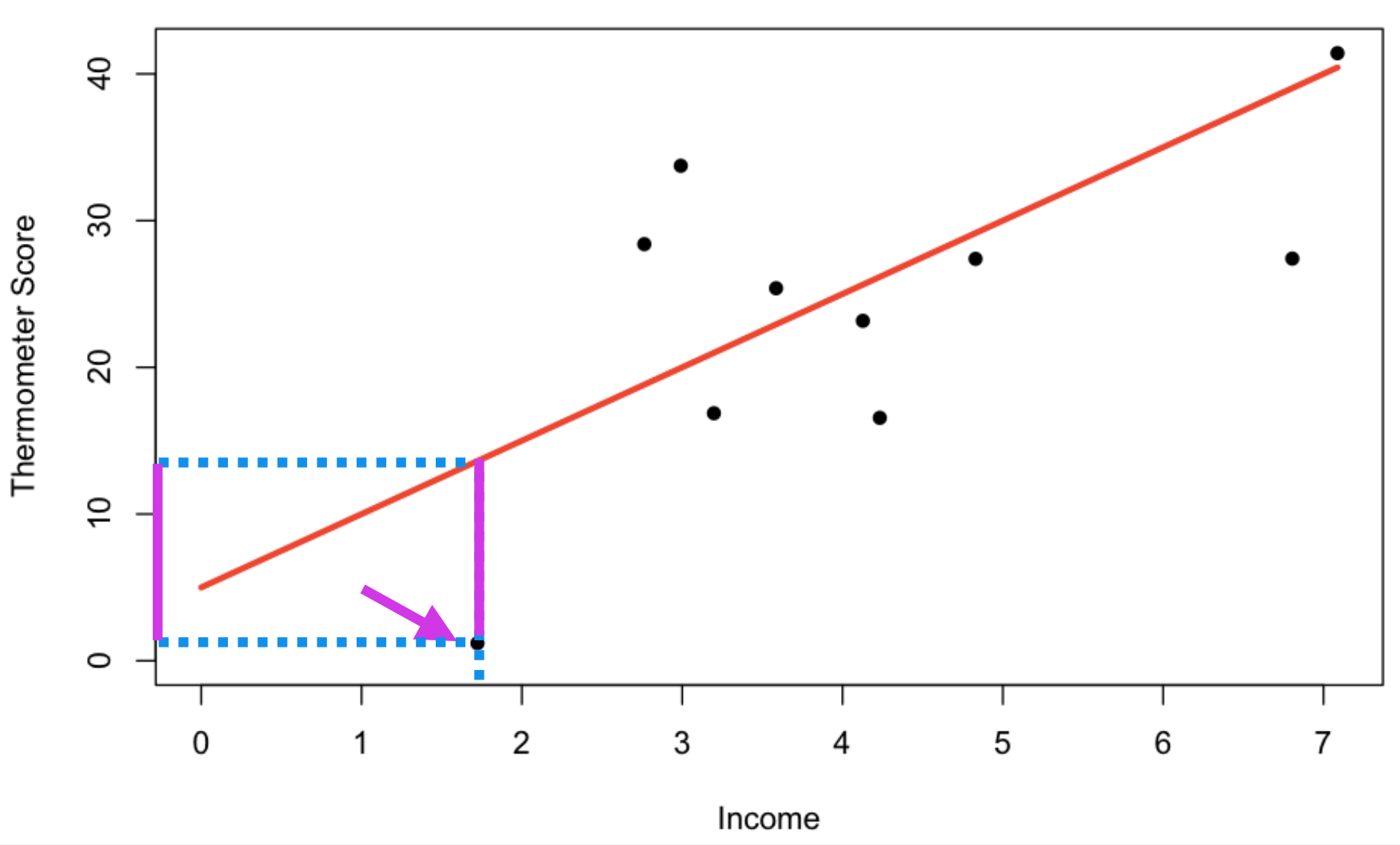
- For each observation, we have a prediction error: $y - \hat{y}$
  - Some are positive, some are negative
- We square the prediction errors: $(y - \hat{y})^2$
  - Now all are positive
  - Squared prediction errors especially large for predictions that are way off
    - e.g. prediction error 2 vs. 20
    - squared prediction errors will be 4 vs. 400

SQUARED PREDICTION ERROR



- Prediction error: $y - \hat{y} = 28 - 19 = 9$
- Squared prediction error: $9^2=81$

SQUARED PREDICTION ERROR

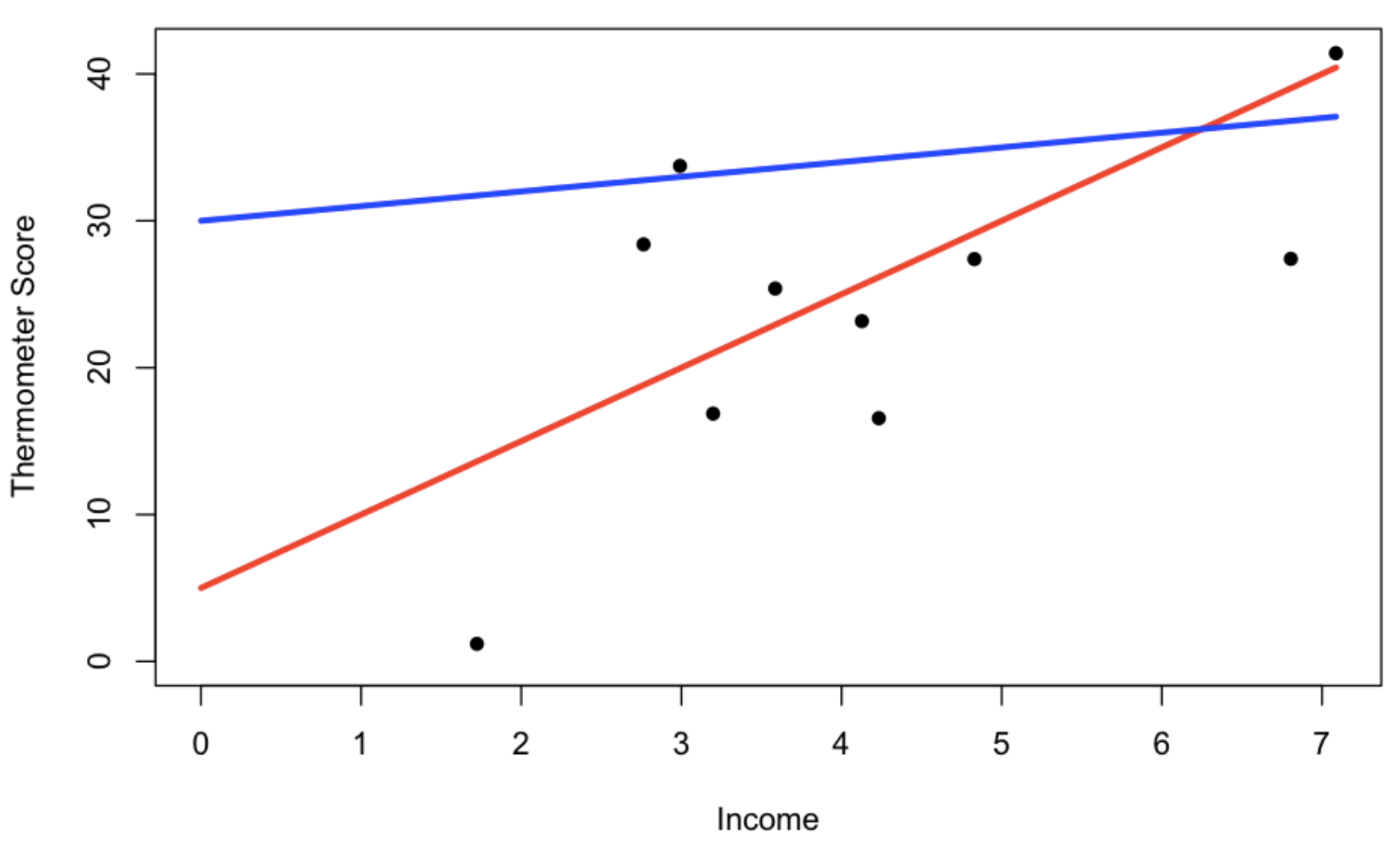


- Prediction error: $y - \hat{y} = 1 - 14 = -13$
- Squared prediction error: $(-13)^2 = 169$

SQUARED PREDICTION ERROR

- We sum squared prediction errors for all observations
- 81 + 169 + all the other observations = 696

SQUARED PREDICTION ERROR

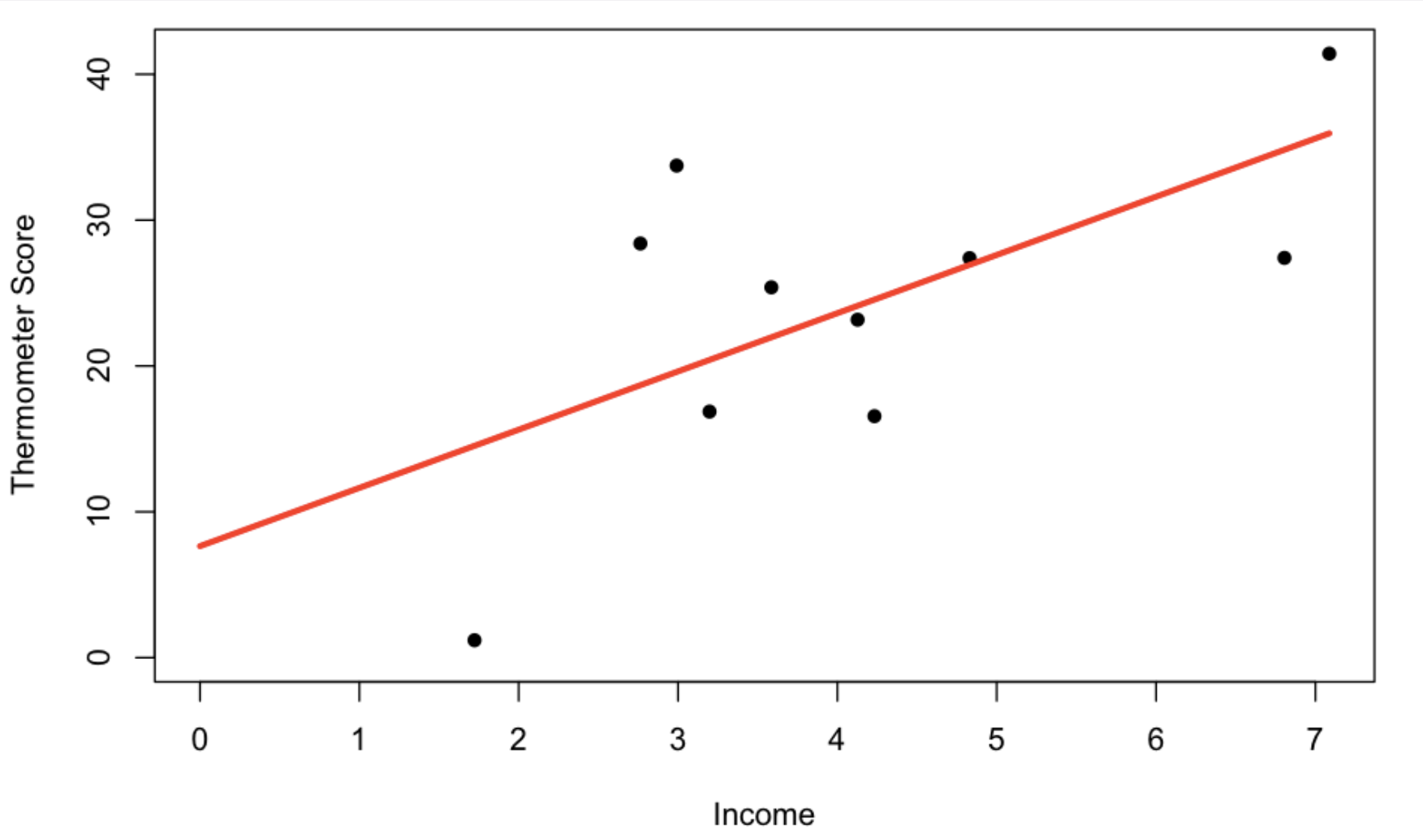


- Sum of squared prediction error red line: 696
- Sum of squared prediction error blue line: 1880

BEST LINE

- The best line is the one with the smallest sum of squared prediction errors
- “Ordinary Least Squares” (OLS) Linear Regression

BEST-FITTING LINE



- **Sum of squared prediction errors: 646.3**

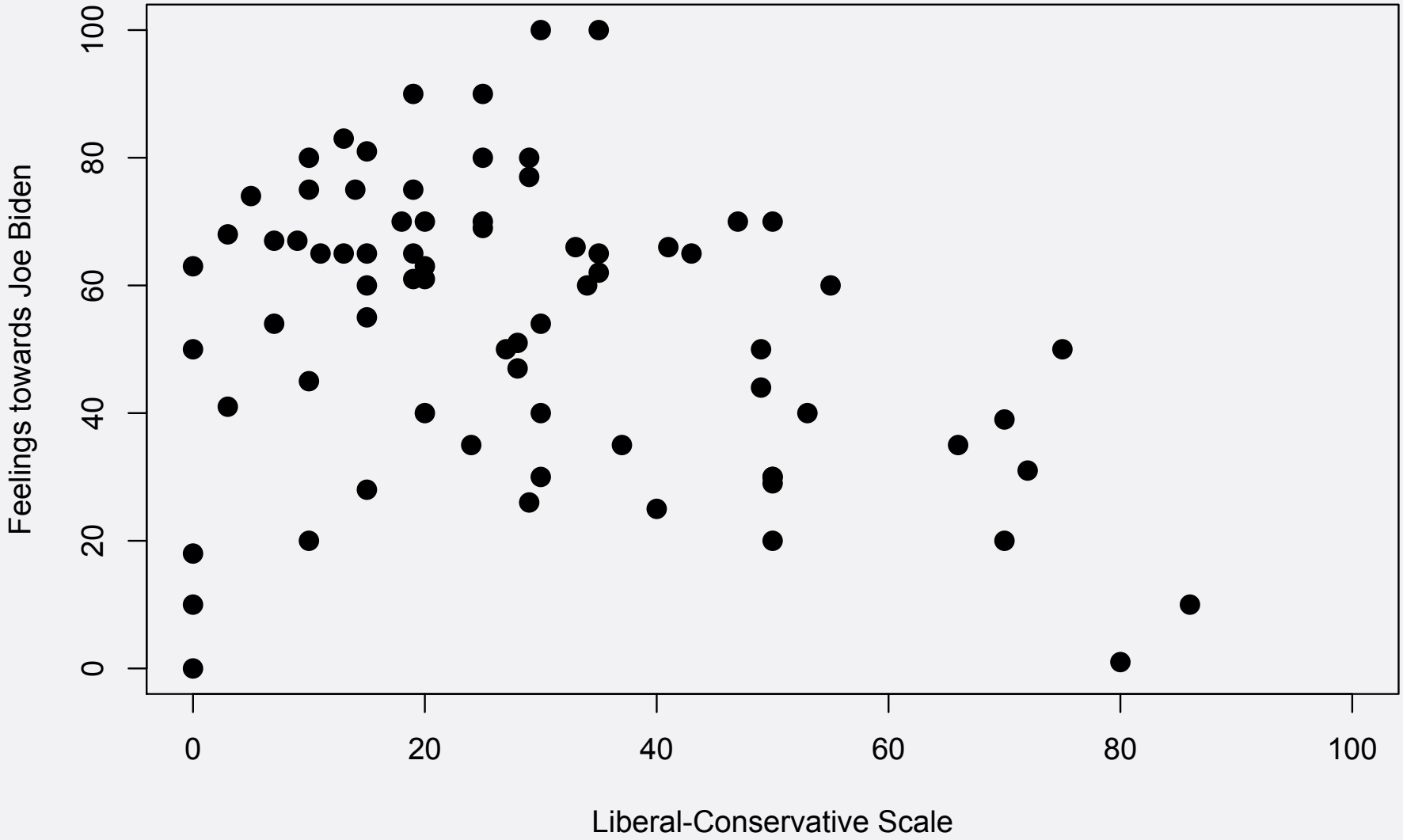
FINDINGS THE BEST LINE

- There is a lot of complicated math behind how to find the best line

$$\hat{\beta} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{\operatorname{Cov}[x,y]}{\operatorname{Var}[x]}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}\,.$$

- Thankfully there are computer programs like R or Stata that do this for us….

BACK TO BIDEN EXAMPLE



BACK TO OUR EXAMPLE

```
. reg therm_2 libcons_1
```

Source	SS	df	MS		
Model	3834.01698	1	3834.01698	Prob > F	0.0059
Residual	34232.5776	72	475.452467	R-squared	0.1007
Total	38066.5946	73	521.4602	Root MSE	21.805

Number of obs	=	74
F(1, 72)	=	8.06
Prob > F	=	0.0059
R-squared	=	0.1007
Adj R-squared	=	0.0882
Root MSE	=	21.805

therm_2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
libcons_1	-.347605	.1224088	-2.84	0.006	-.5916224	-.1035876
_cons	63.79618	4.3579	14.64	0.000	55.10887	72.4835

- DV: Rating of J. Biden (therm_2)
- IV: Liberal-conservative scale (libcons_1)

BACK TO OUR EXAMPLE

```
. reg therm_2 libcons_1
```

Source	SS	df	MS		
Model	3834.01698	1	3834.01698	Prob > F	0.0059
Residual	34232.5776	72	475.452467	R-squared	0.1007
Total	38066.5946	73	521.4602	Root MSE	21.805

Number of obs	=	74
F(1, 72)	=	8.06
Prob > F	=	0.0059
R-squared	=	0.1007
Adj R-squared	=	0.0882
Root MSE	=	21.805

therm_2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
libcons_1	-.347605	.1224088	-2.84	0.006	-.5916224	-.1035876
_cons	63.79618	4.3579	14.64	0.000	55.10887	72.4835

Intercept

BACK TO OUR EXAMPLE

```
. reg therm_2 libcons_1
```

Source	SS	df	MS		
Model	3834.01698	1	3834.01698	Prob > F	0.0059
Residual	34232.5776	72	475.452467	R-squared	0.1007
Total	38066.5946	73	521.4602	Root MSE	21.805

Number of obs	=	74
F(1, 72)	=	8.06
Prob > F	=	0.0059
R-squared	=	0.1007
Adj R-squared	=	0.0882
Root MSE	=	21.805

therm_2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
libcons_1	-.347605	.1224088	-2.84	0.006	-.5916224	-.1035876
_cons	63.79618	4.3579	14.64	0.000	55.10887	72.4835

Slope

BACK TO OUR EXAMPLE

```
. reg therm_2 libcons_1
```

Source	SS	df	MS		
Model	3834.01698	1	3834.01698	Number of obs	= 74
Residual	34232.5776	72	475.452467	F(1, 72)	= 8.06
				Prob > F	= 0.0059
				R-squared	= 0.1007
				Adj R-squared	= 0.0882
Total	38066.5946	73	521.4602	Root MSE	= 21.805

therm_2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
libcons_1	-.347605	.1224088	-2.84	0.006	-.5916224	-.1035876
_cons	63.79618	4.3579	14.64	0.000	55.10887	72.4835

- Thermometer Score = 63.80 - 0.348 * Lib/Cons
- (I simplified numbers earlier to make math easier…)