PSC 202
SYRACUSE UNIVERSITY

INTRODUCTION TO POLITICAL ANALYSIS

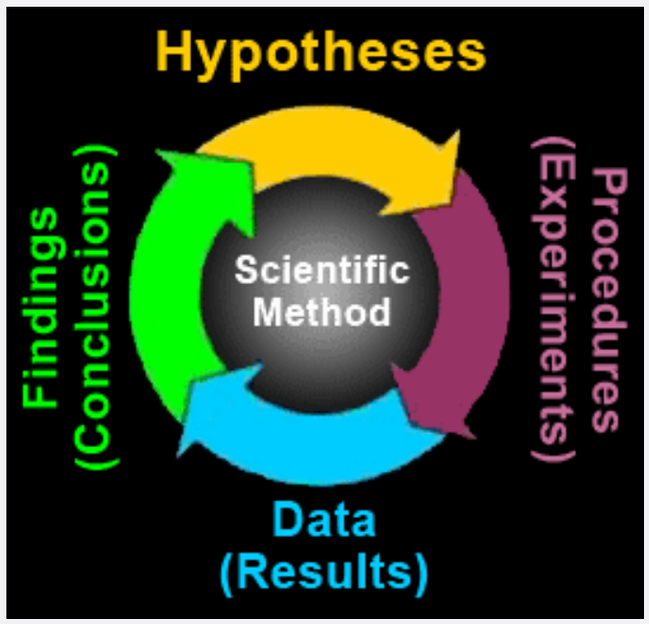
BIVARIATE HYPOTHESIS TESTING PART 3

REMINDER: EXAM #2

- Originally scheduled for March 27 (next Monday)
- Moved to April 3 (one week later)

WHERE WE ARE

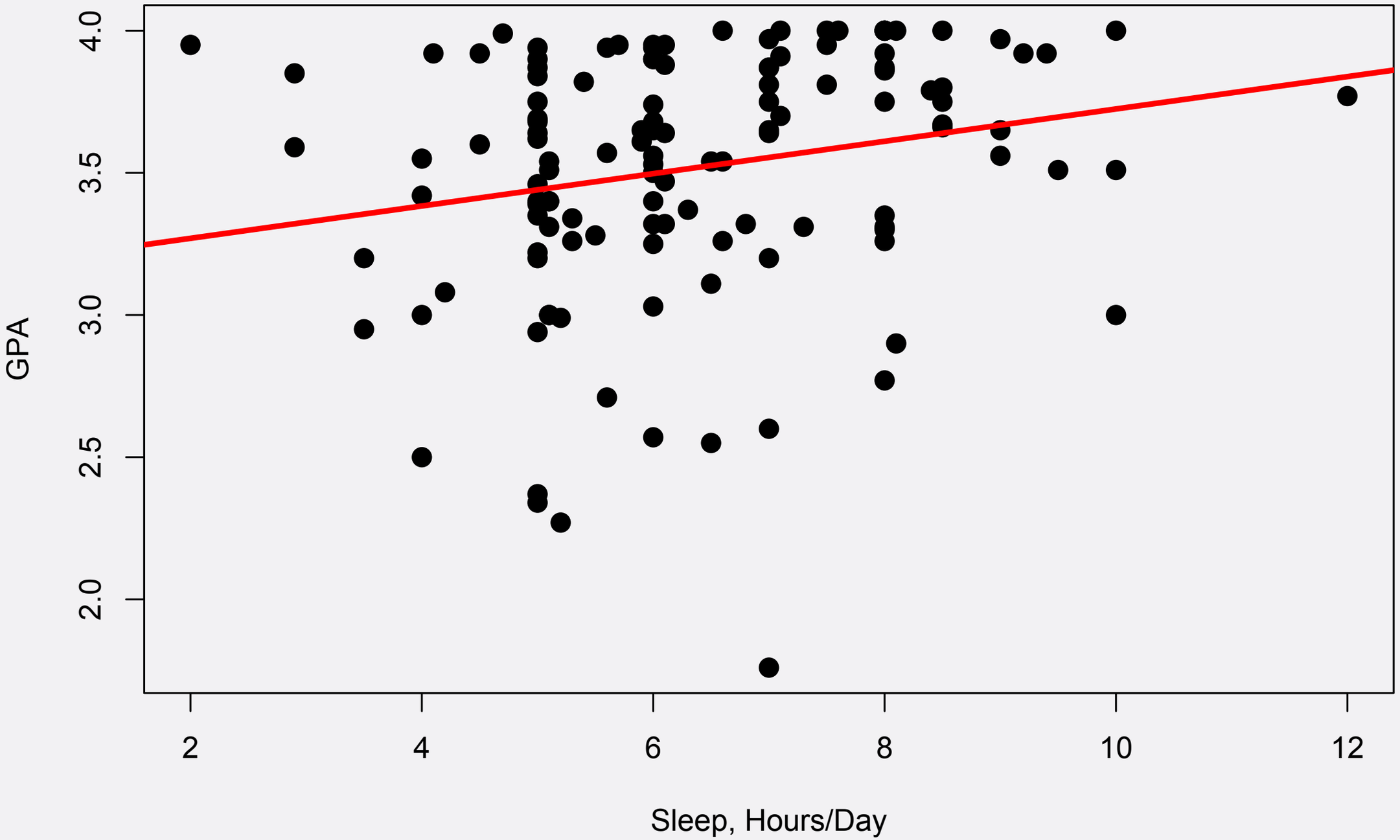
- Formulate research question
- Propose explanation/theory, hypotheses
- Data collection process
- Use data to evaluate hypotheses
- Reassess explanation



BIVARIATE RELATIONSHIPS

Dependent Variable	Independent Variable: Nominal/Ordinal	Independent Variable: Interval
Nominal/Ordinal	Cross-Tabulation	Not In This Class…
Interval	Mean Comparison	Correlation Coefficient, Linear Regression

LINEAR REGRESSION



- GPA = 3.2 + 0.06 * Hours of Sleep

INTERPRETATION?

- GPA = 3.2 + 0.06 * Hours of Sleep
- **What does the 3.2 tell us?**
- **What does the 0.06 tell us?**

INTERPRETATION

- GPA = 3.2 + 0.06 * Hours of Sleep
  - What does the 3.2 tell us?
    - A student who sleeps 0 hours per day has an expected GPA of 3.2
  - What does the 0.06 tell us?
    - For every additional hour of sleep per night, GPA is expected to increase by 0.06 points

WHAT THIS TELLS US

- GPA = 3.2 + 0.06 * Hours of Sleep
- Expected GPA of someone sleeping 4 hours per night
  - 3.2 + 0.06 * 4 = 3.44

WHAT THIS TELLS US

- GPA = 3.2 + 0.06 * Hours of Sleep
- Expected GPA of someone sleeping 8 hours per night
  - 3.2 + 0.06 * 8 = 3.68

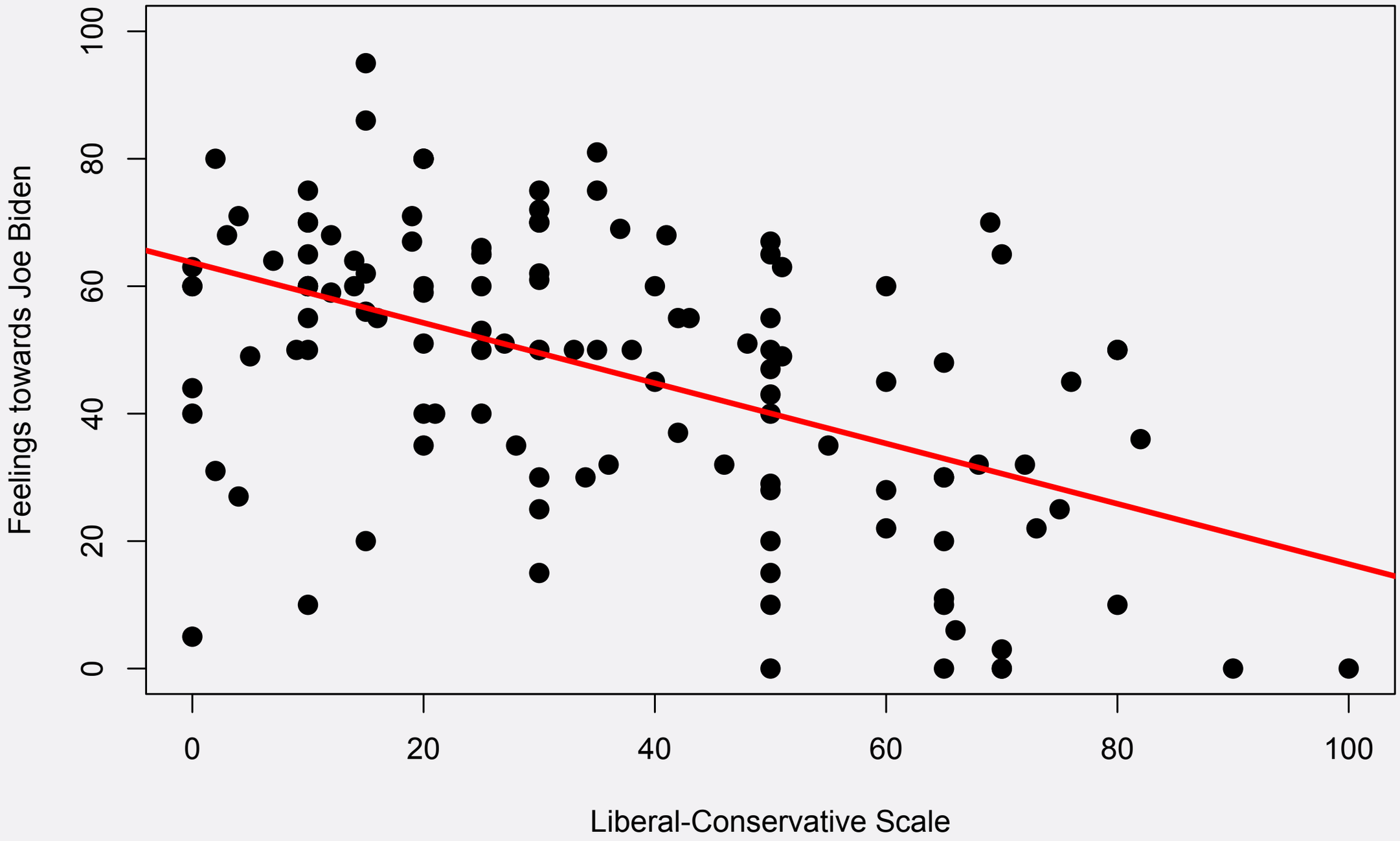
LINEAR REGRESSION

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

TODAY

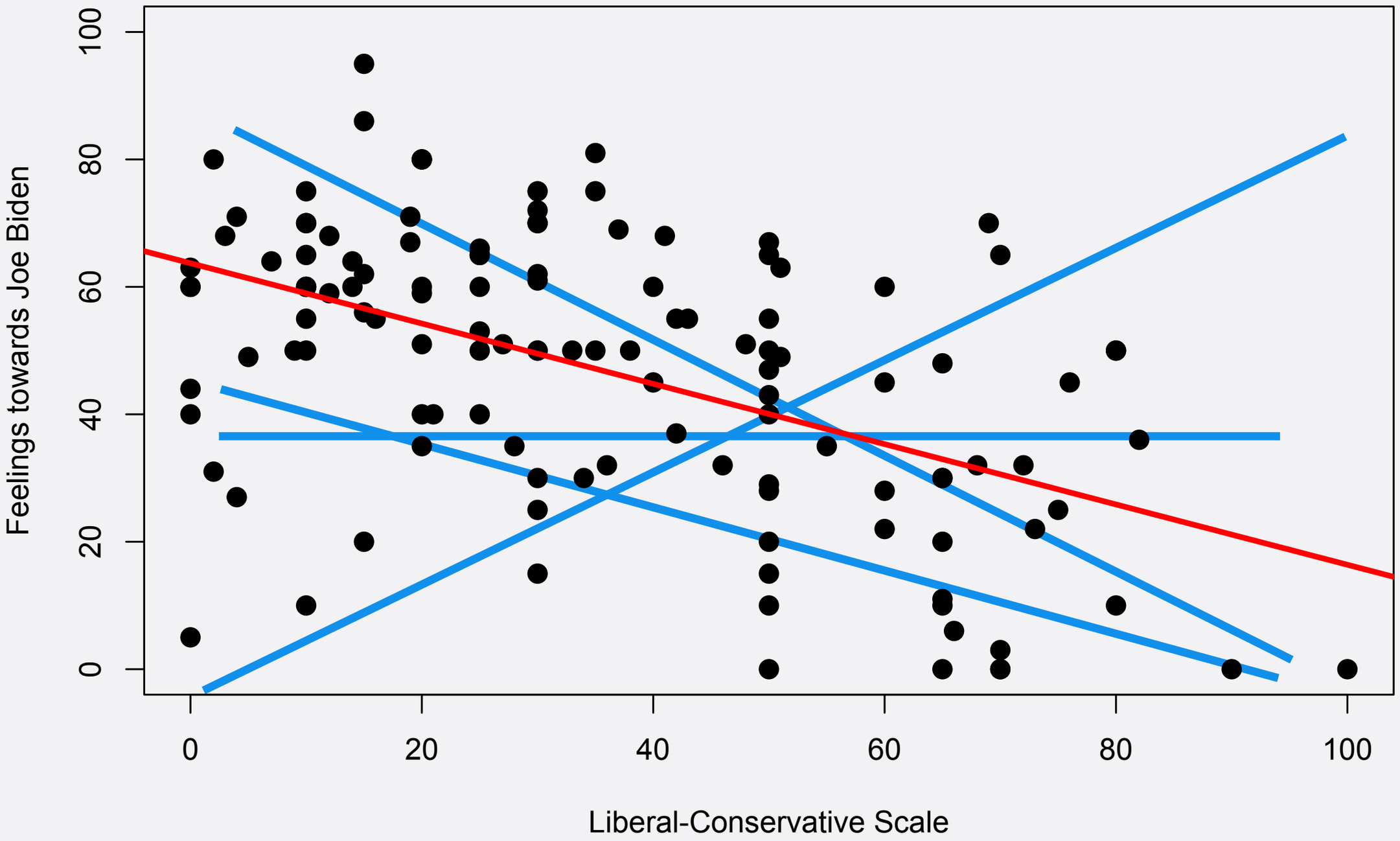
- How do I pick the line?
- How is linear regression useful?
- Caveats about linear regression

HOW TO PICK THE LINE



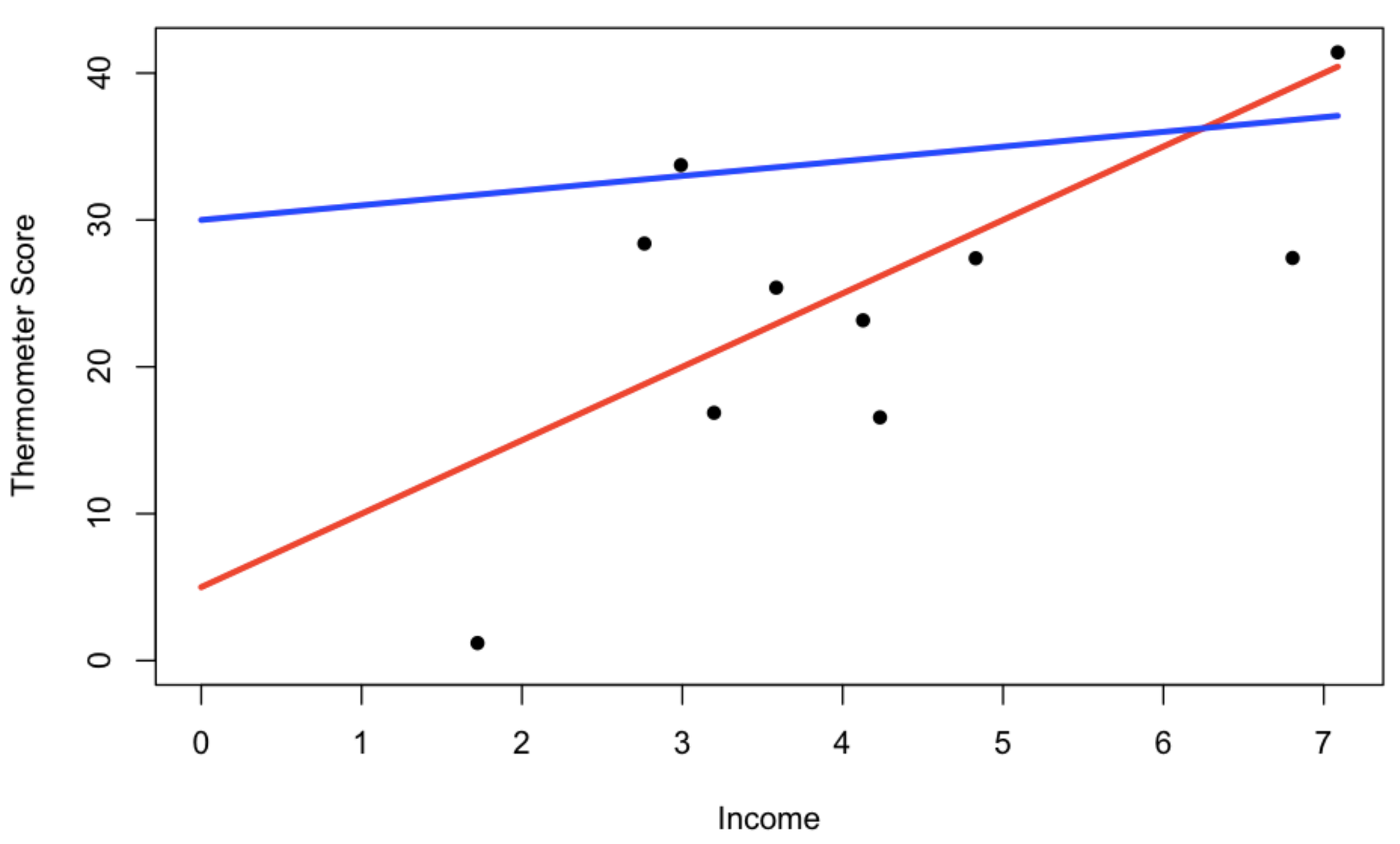
- **Why this line?**

HOW TO PICK THE LINE



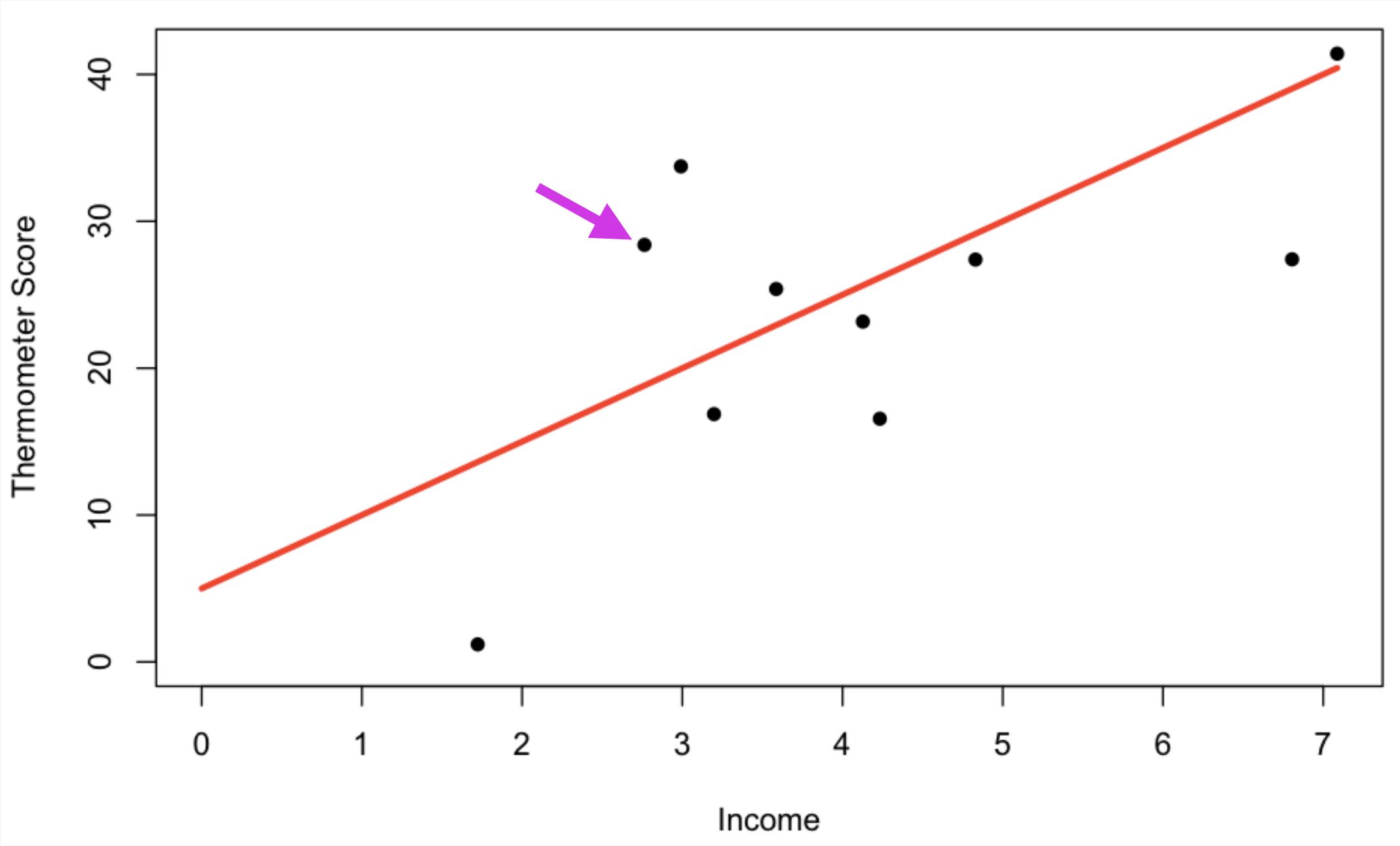
- **Why not any of these?**

HOW TO PICK THE LINE

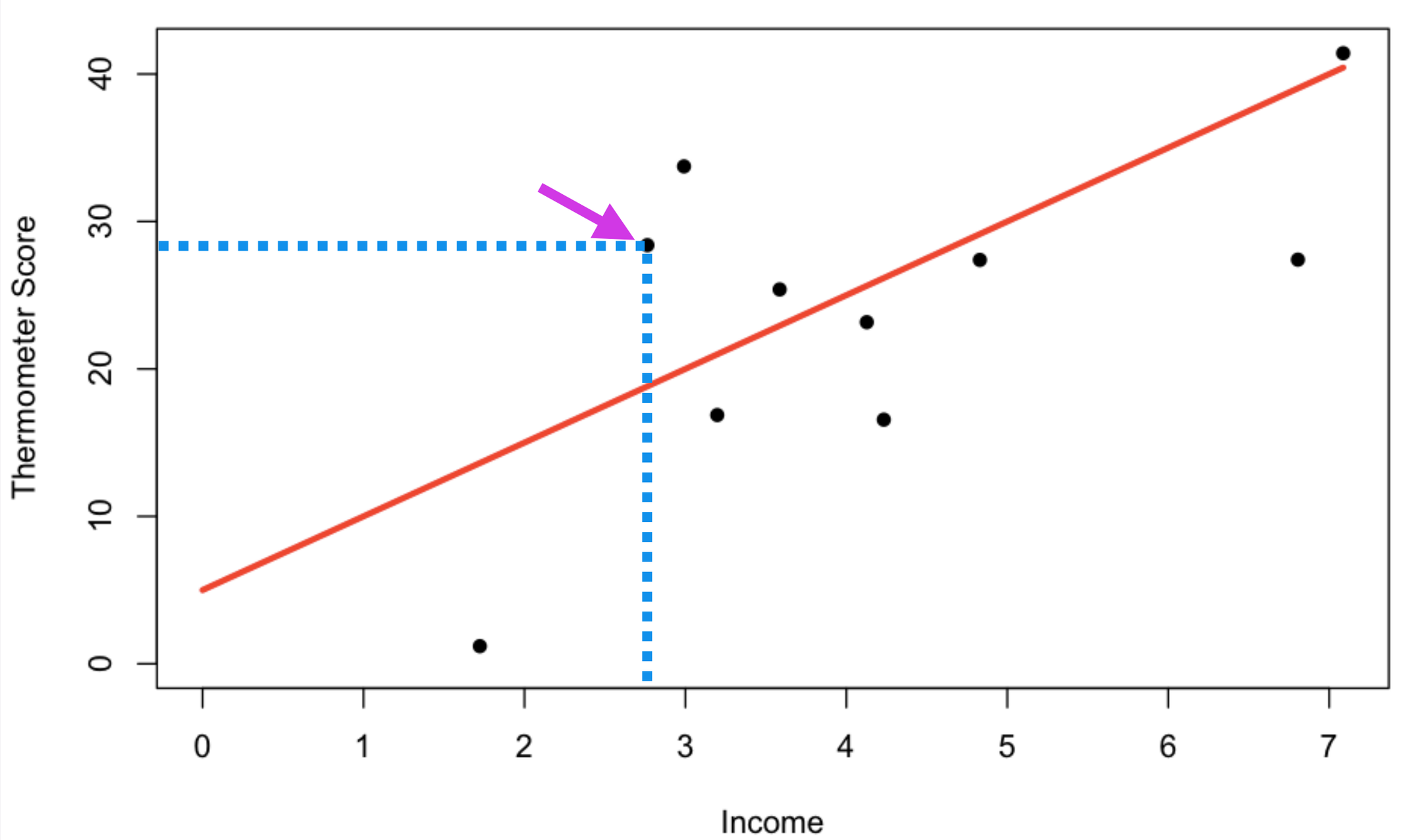


- **Which line is better?**

HOW TO PICK THE LINE

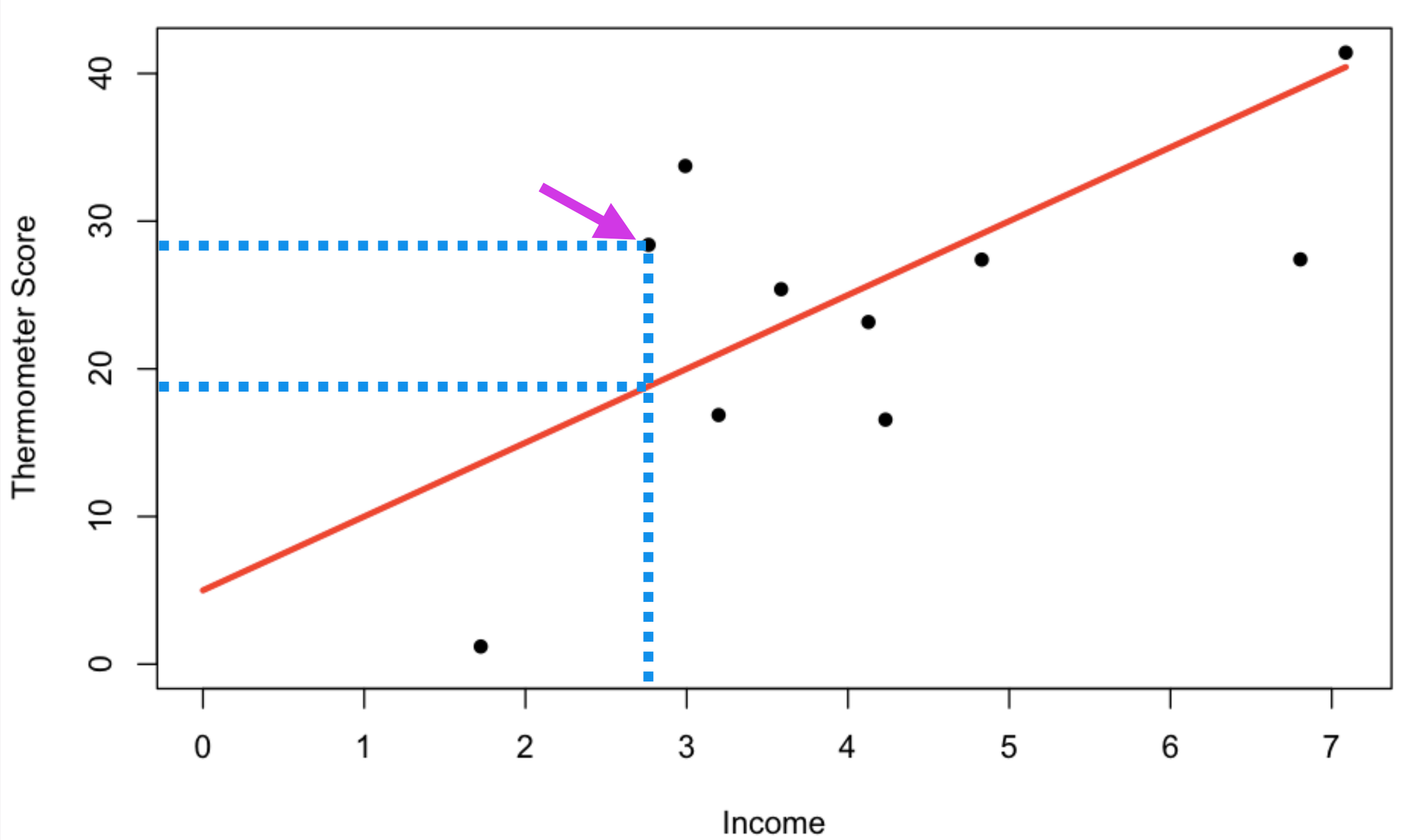


HOW TO PICK THE LINE



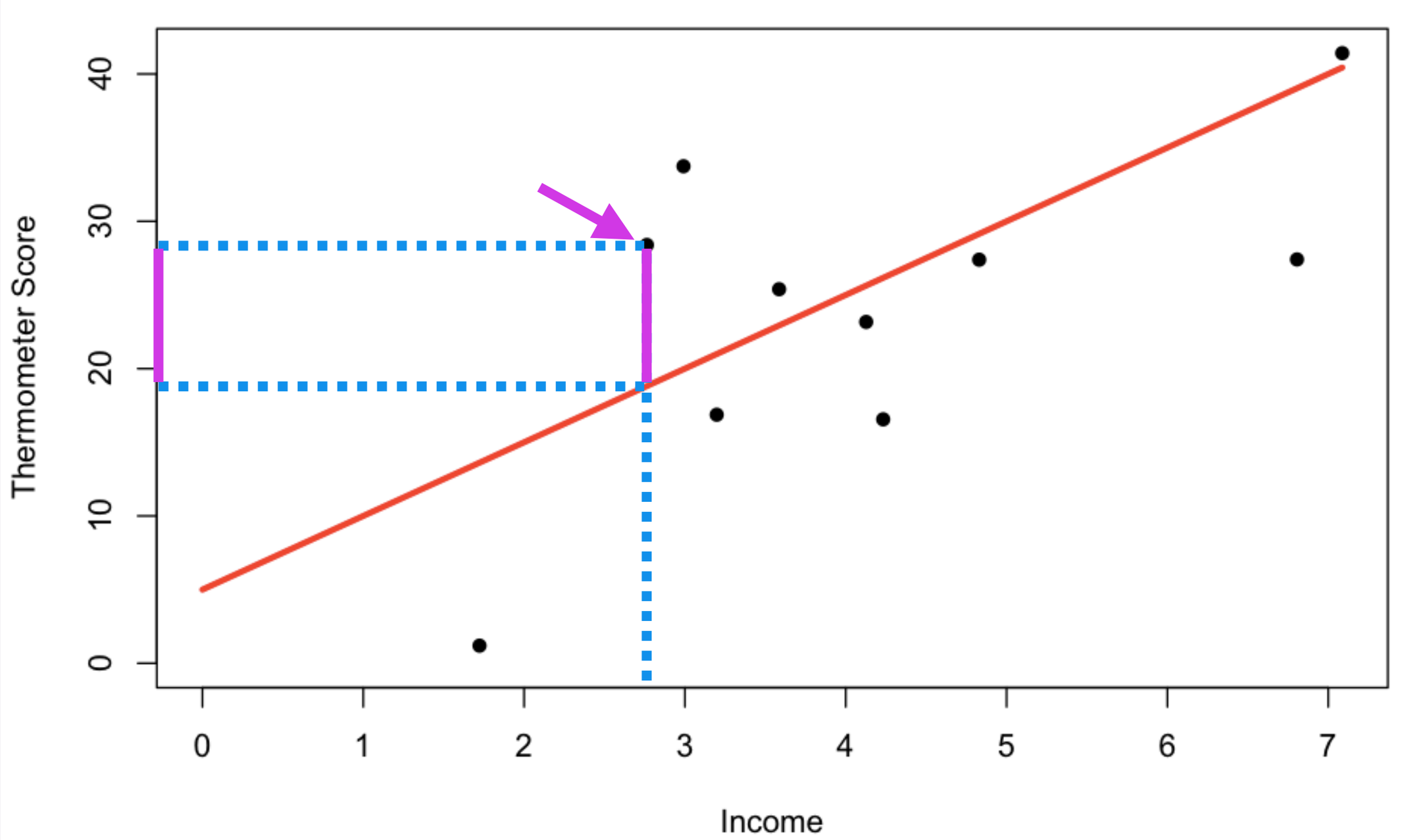
- **Actual y-value: y=28**

HOW TO PICK THE LINE



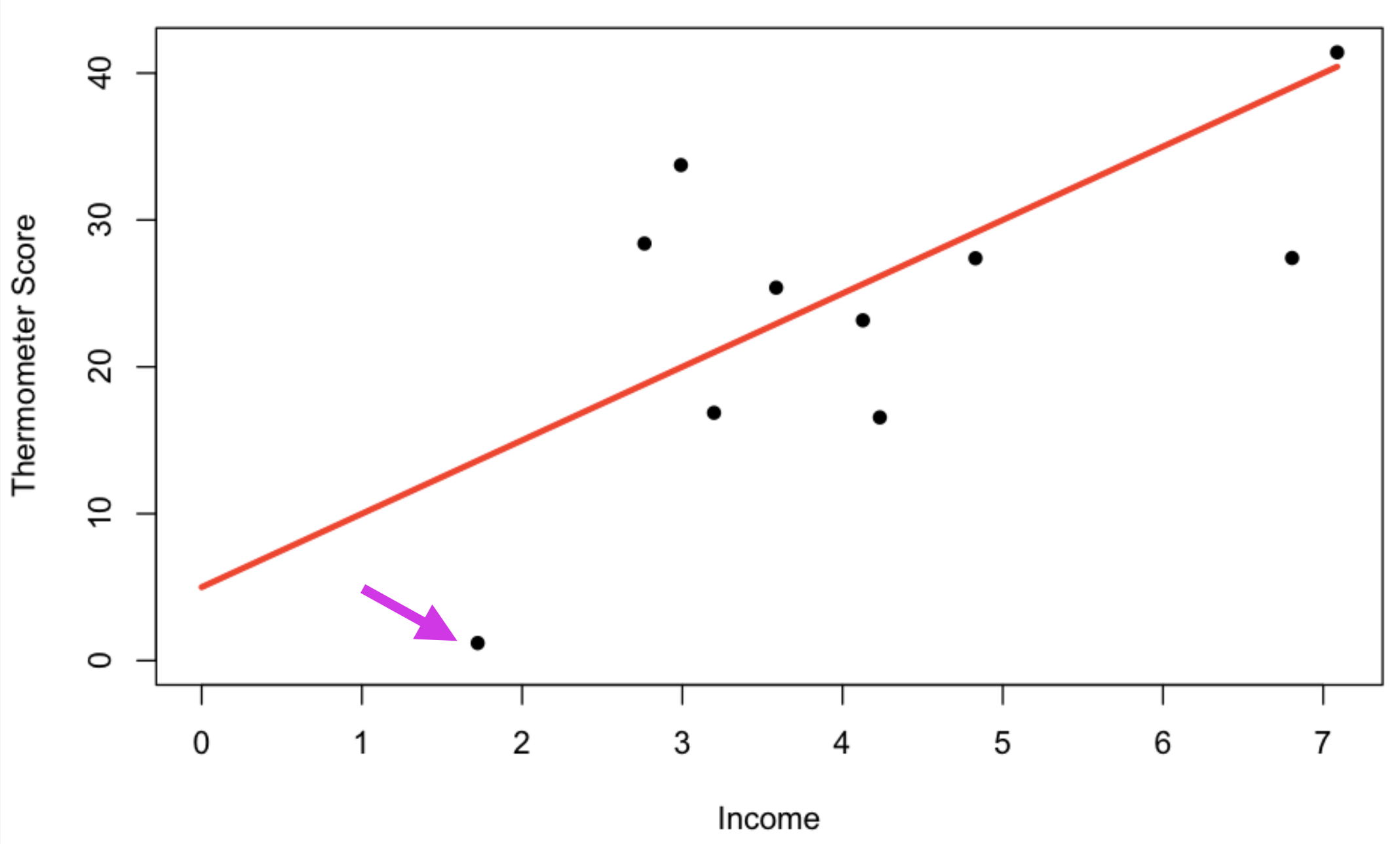
- **Predicted y-value: $\hat{y}=19$**

HOW TO PICK THE LINE

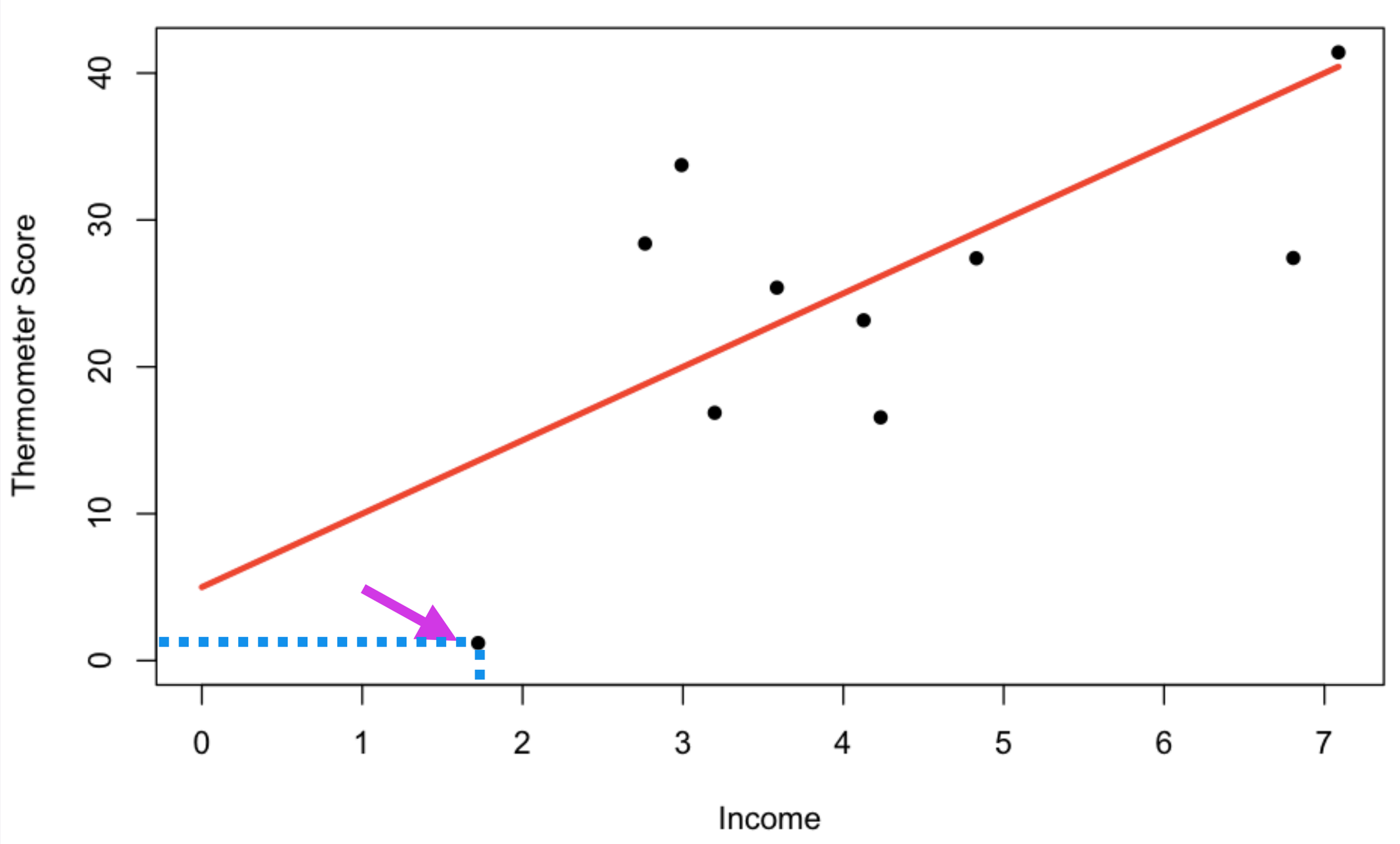


- Prediction error: $y - \hat{y} = 28 - 19 = 9$

HOW TO PICK THE LINE

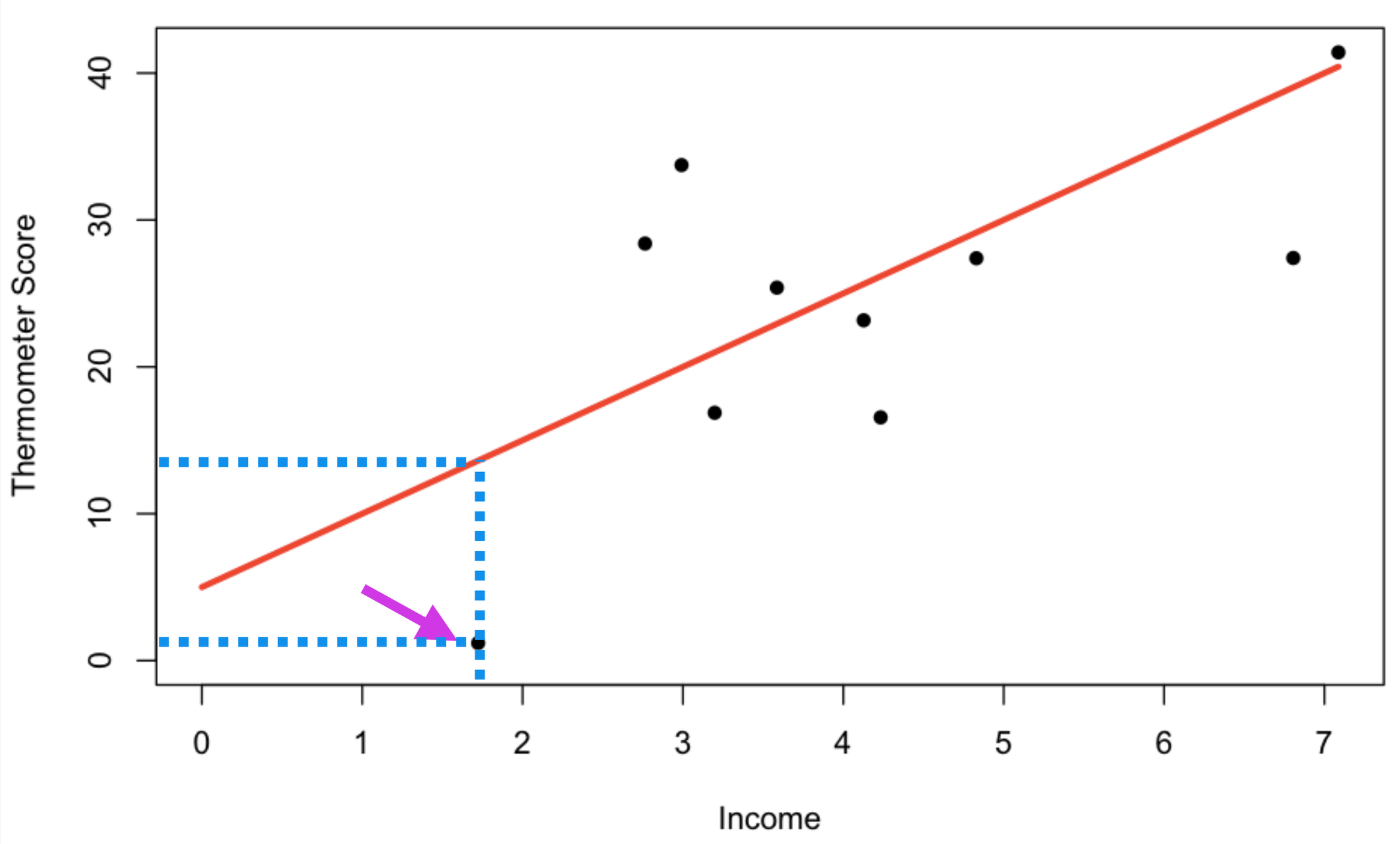


HOW TO PICK THE LINE



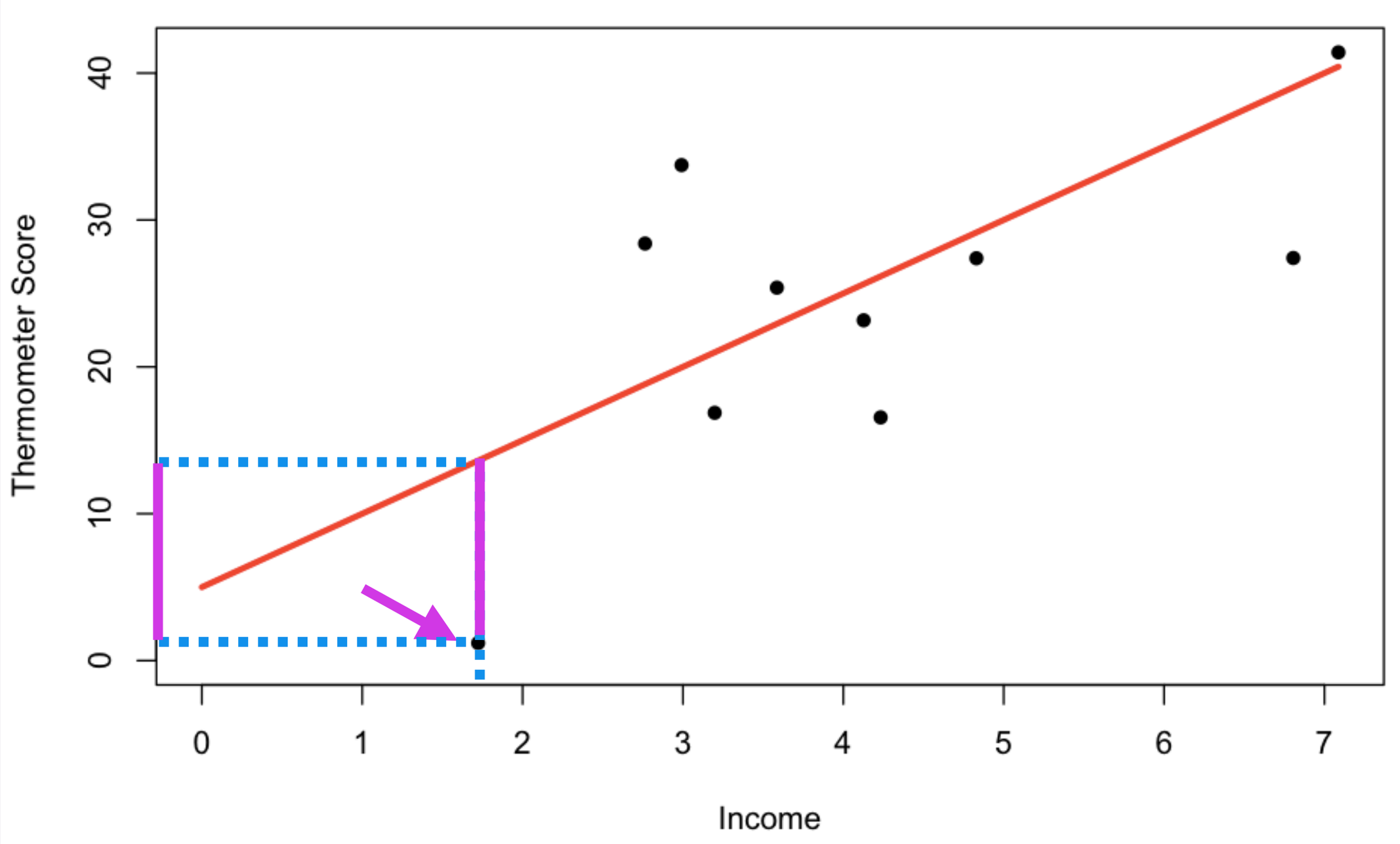
- **Actual y-value: y=1**

HOW TO PICK THE LINE



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

HOW TO PICK THE LINE

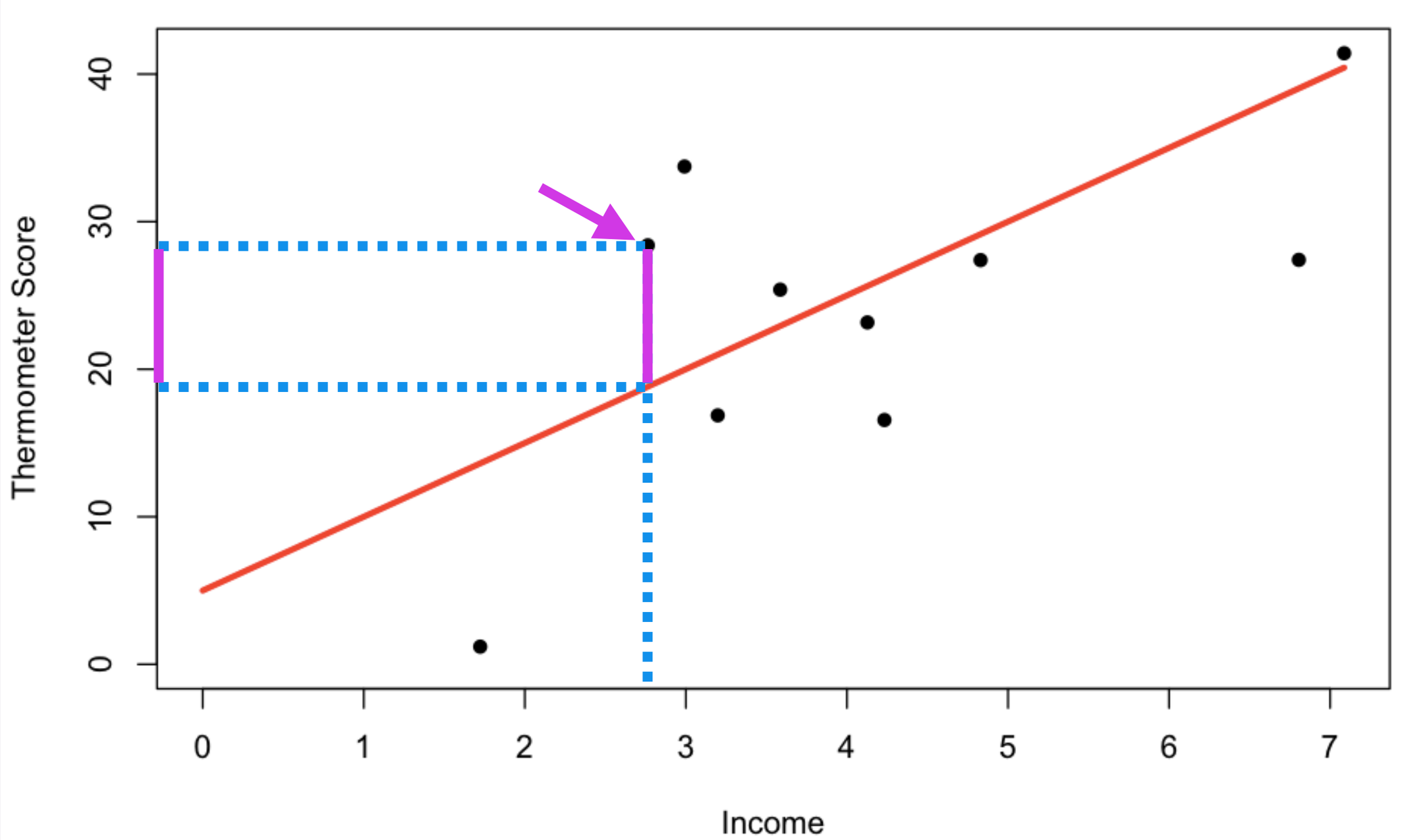


- Prediction error: $y - \hat{y} = 1 - 14 = -13$

PREDICTION ERROR

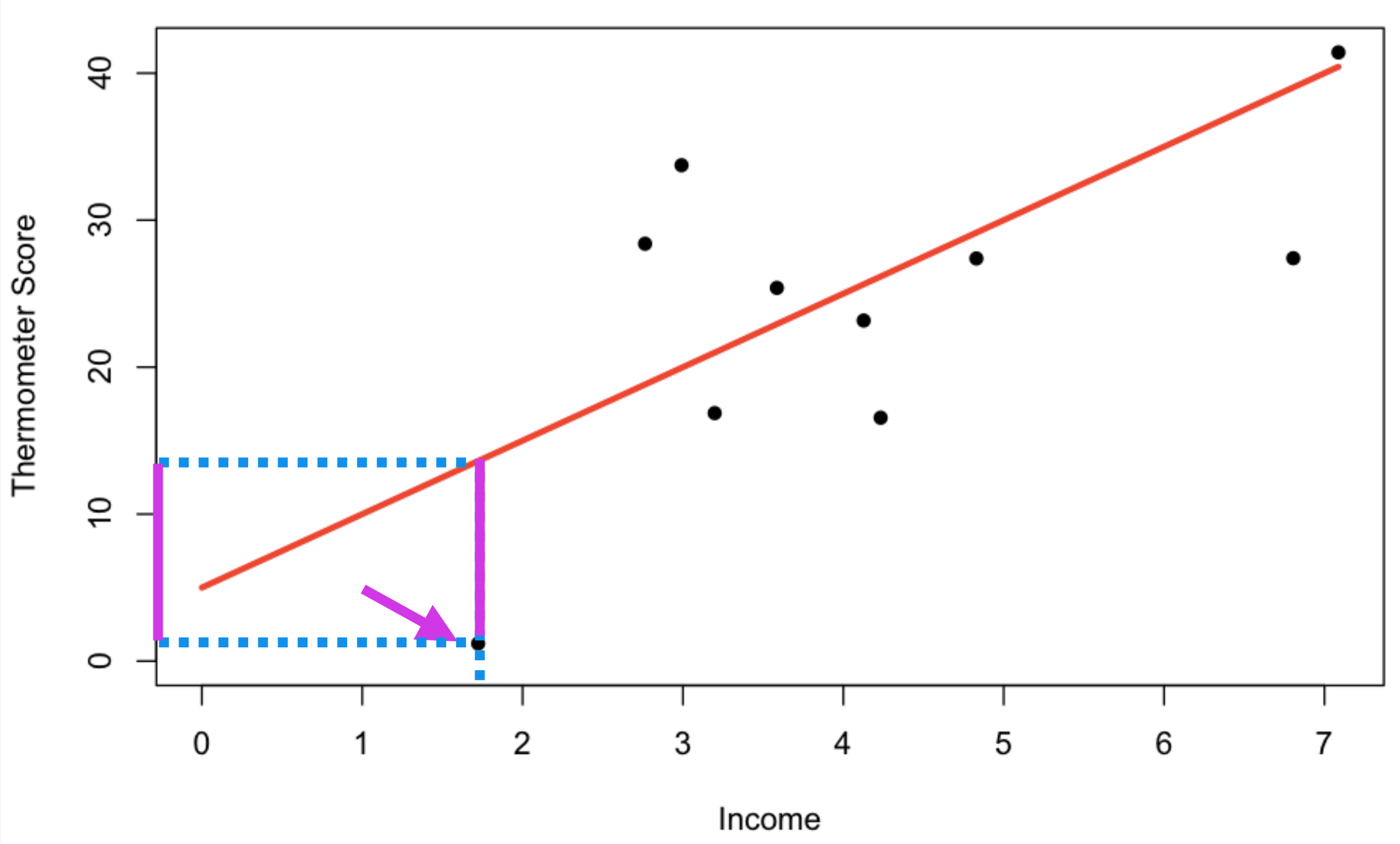
- For each observation, we have a prediction error: $y - \hat{y}$
  - Some are positive, some are negative
- We square the prediction errors: $(y - \hat{y})^2$
  - Now all are positive

SQUARED PREDICTION ERROR



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

SQUARED PREDICTION ERROR

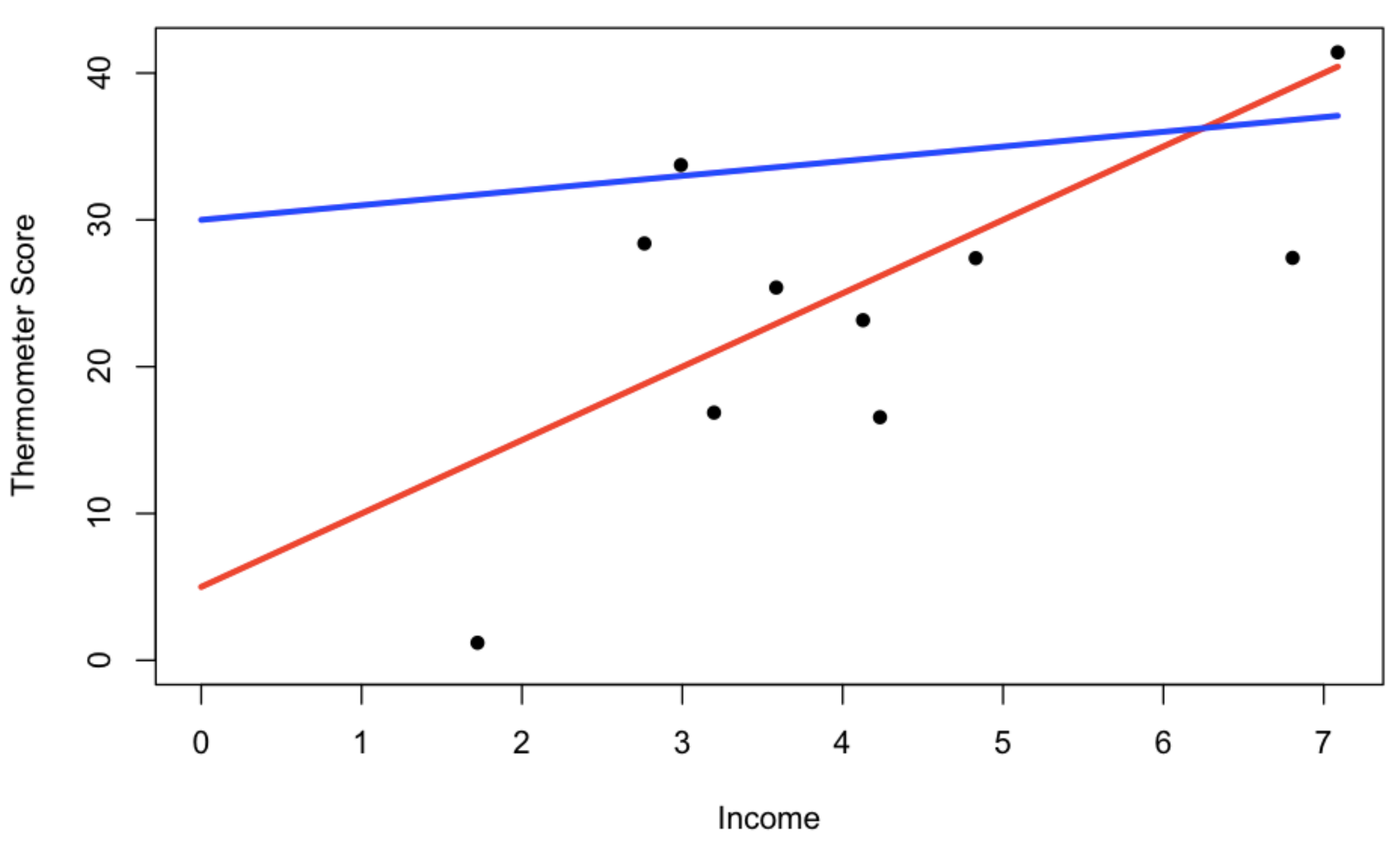


- Prediction error: $y - \hat{y} = 1 - 14 = -13$
- Squared prediction error: $(-13)^2 = 169$

SQUARED PREDICTION ERROR

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

SQUARED PREDICTION ERROR

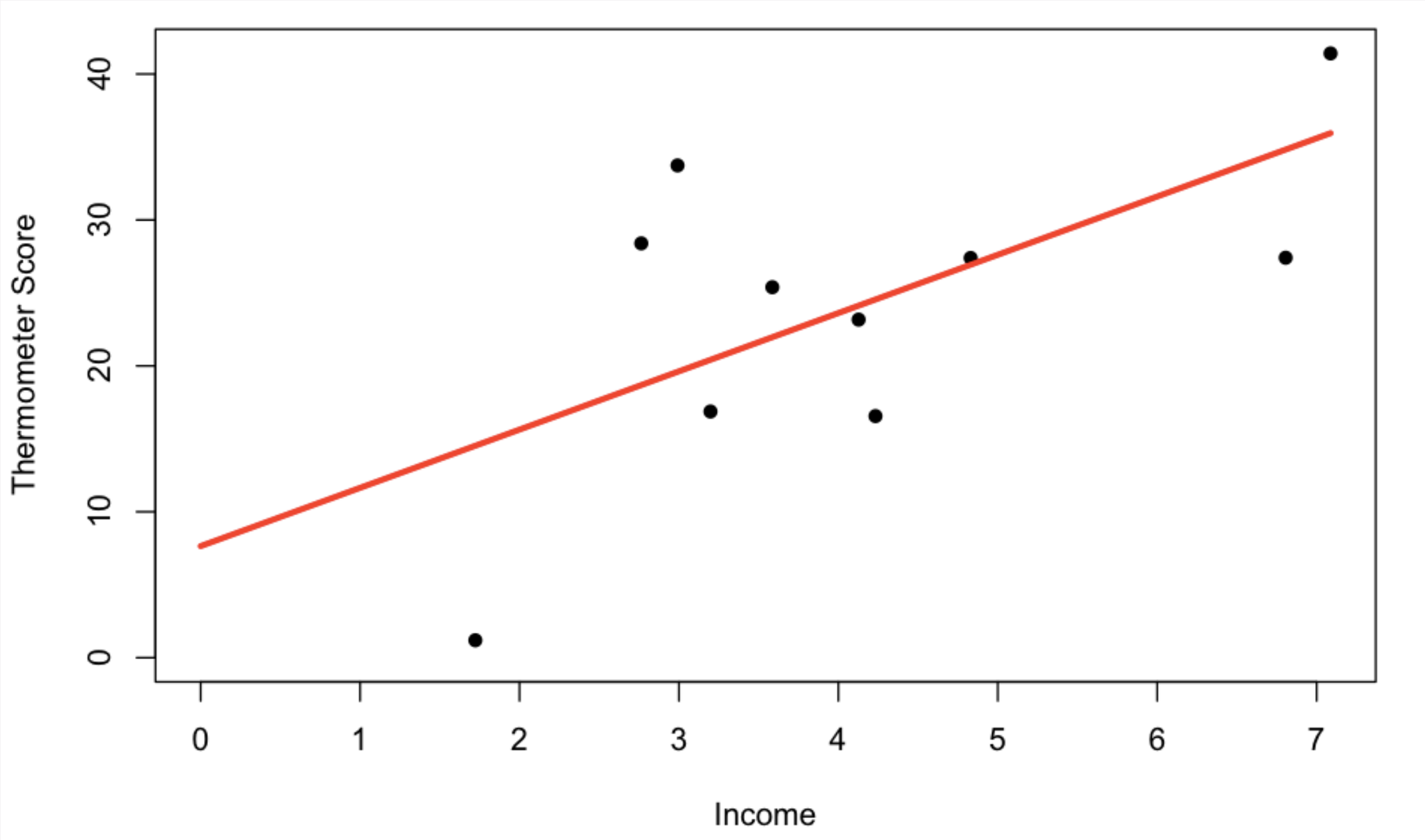


- Sum of squared prediction error red line: 696
- Sum of squared prediction error blue line: 1880

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**

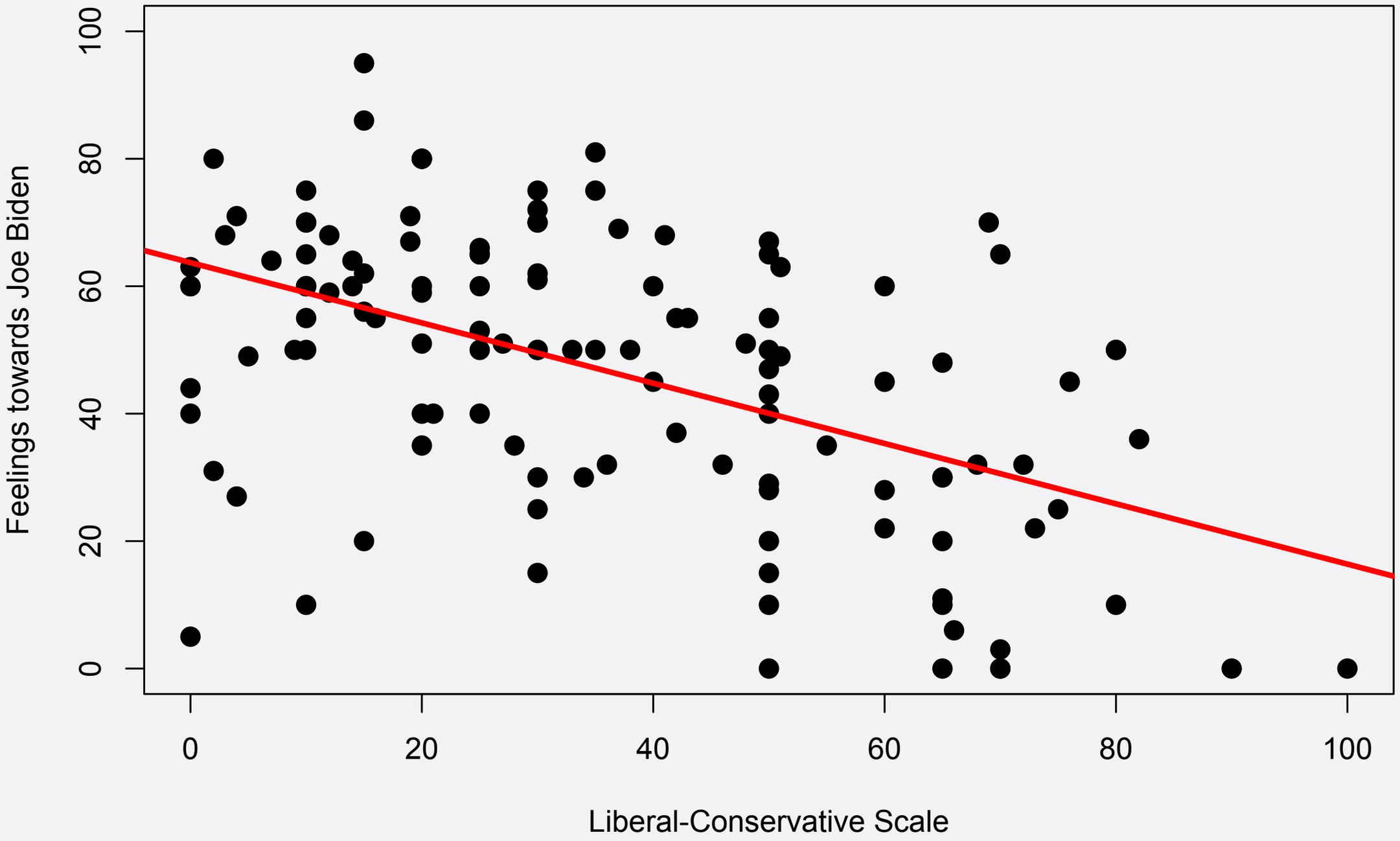
FINDINGS THE BEST LINE

- 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

```
> summary(lm(therm_2 ~ libcons_1, data = data))

Call:
lm(formula = therm_2 ~ libcons_1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-58.713 -12.954   1.019  12.484  38.939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.71327    3.09127  20.611  < 2e-16 ***
libcons_1   -0.47323    0.07174  -6.597 1.12e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.23 on 123 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.2613, Adjusted R-squared:  0.2553
F-statistic: 43.51 on 1 and 123 DF,  p-value: 1.117e-09
```

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

BACK TO OUR EXAMPLE

Intercept

```
> summary(lm(therm_2 ~ libcons_1, data = data))

Call:
lm(formula = therm_2 ~ libcons_1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-58.713 -12.954   1.019  12.484  38.939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.71327    3.09127  20.611  < 2e-16 ***
libcons_1   -0.47323    0.07174  -6.597 1.12e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.23 on 123 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.2613, Adjusted R-squared:  0.2553
F-statistic: 43.51 on 1 and 123 DF,  p-value: 1.117e-09
```

BACK TO OUR EXAMPLE

Slope

```
> summary(lm(therm_2 ~ libcons_1, data = data))

Call:
lm(formula = therm_2 ~ libcons_1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-58.713 -12.954   1.019  12.484  38.939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.71327    3.09127  20.611  < 2e-16 ***
libcons_1   -0.47323    0.07174  -6.597 1.12e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.23 on 123 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.2613, Adjusted R-squared:  0.2553
F-statistic: 43.51 on 1 and 123 DF,  p-value: 1.117e-09
```

- Thermometer Score = 63.71 - 0.47 * Lib/Cons
- (I simplified numbers earlier to make math easier…)

REGRESSION EQUATION

- Thermometer score = 63.71 - 0.47 * Lib/Cons
- General form: y = a + b * x
- y: dependent variable
- a: intercept
- b: slope
- x: independent variable

SLOPE

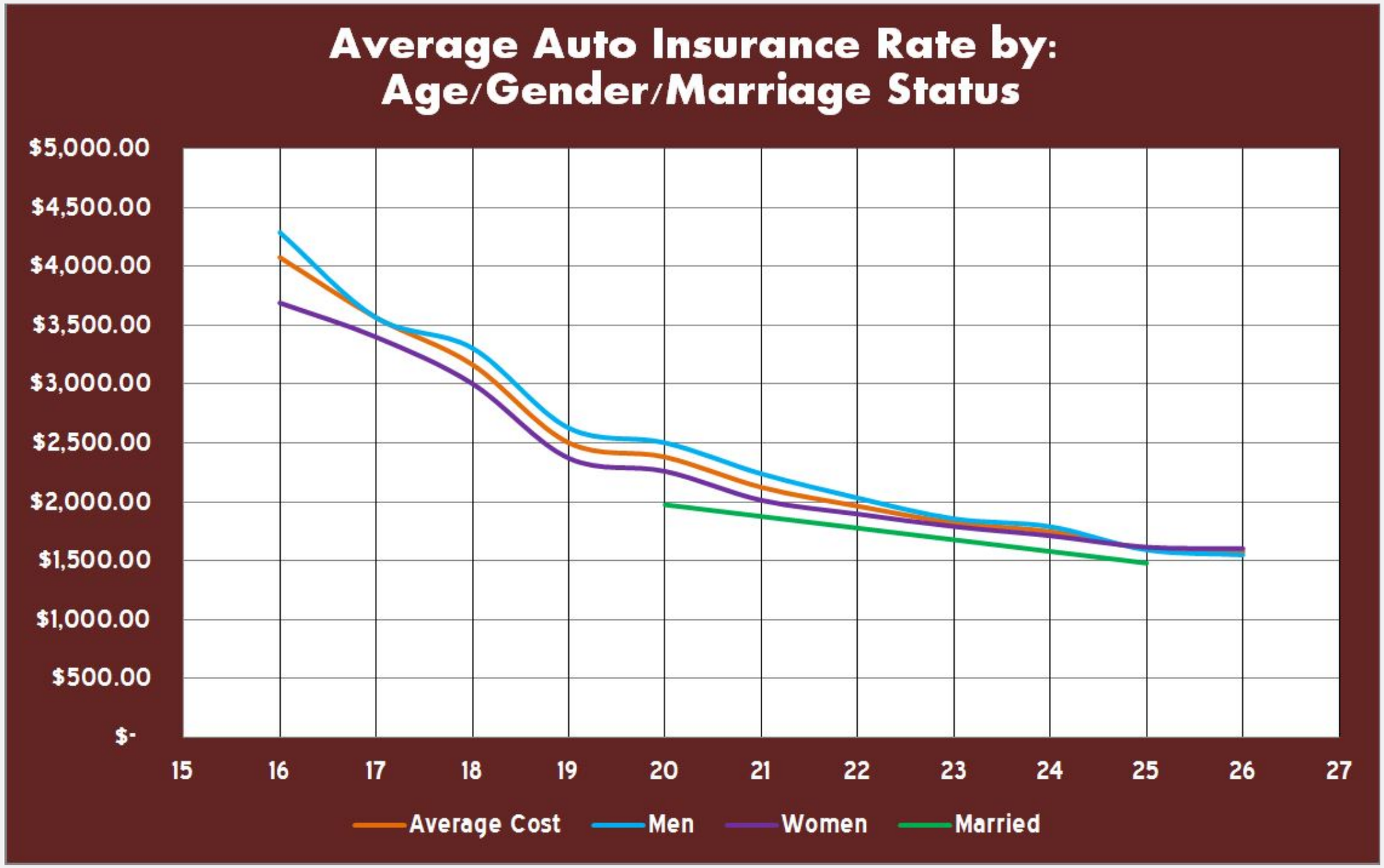
- y = a + b * x
  - Interpretation of slope: For every one unit increase in x, y changes by b units
  - Interpretation of intercept: When x=0, y takes the value a

TODAY

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

HOW IS THIS USEFUL?

- Linear regression widely used



HOW IS THIS USEFUL?

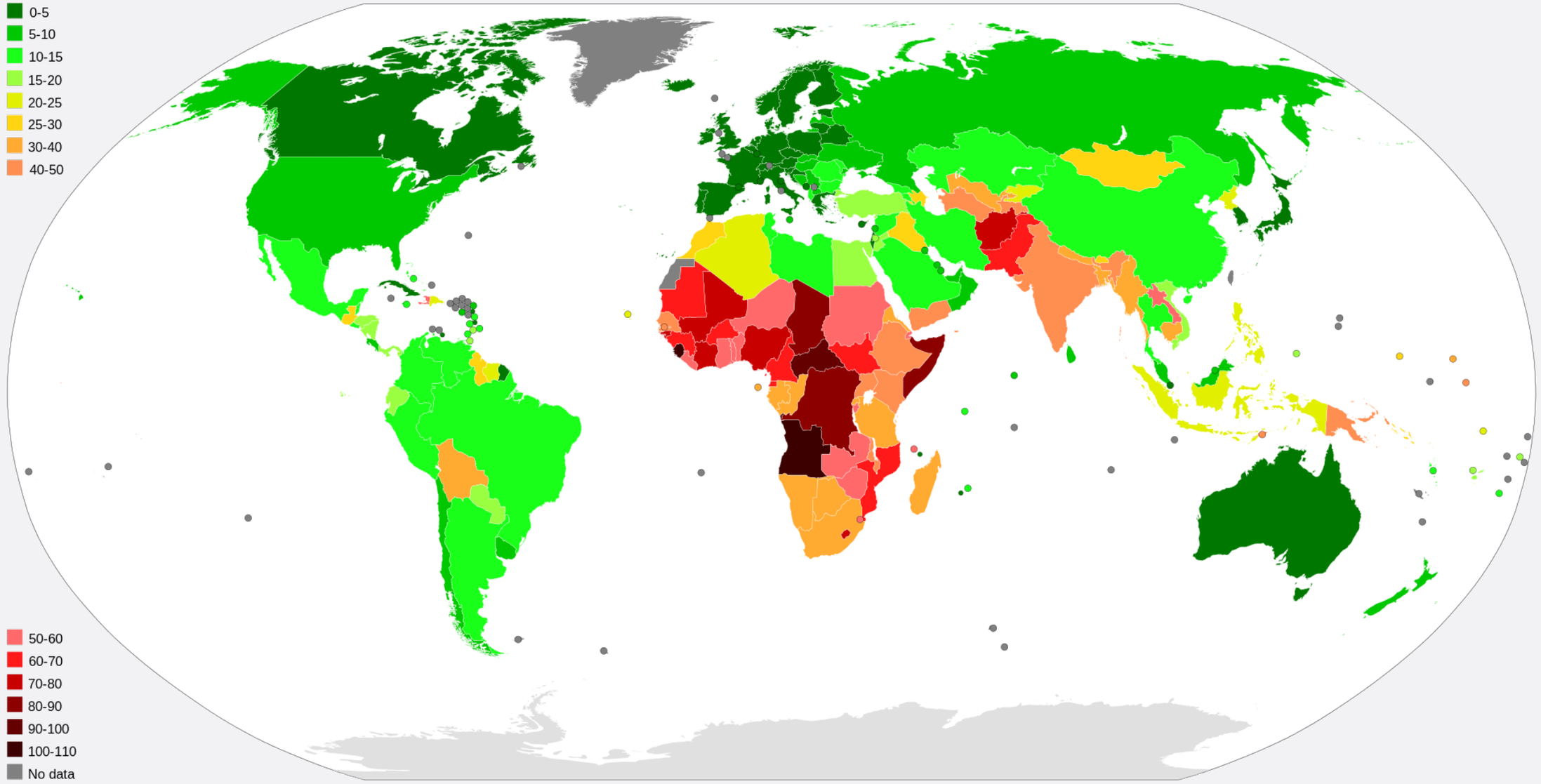
- Insurance company has to decide how much to charge you
- How much to charge you depends on how much in damages they expect to have to pay for you
- Guessing won’t do
  - If they overestimate how much damage someone will cause, they charge too much (and the person might buy insurance elsewhere)
  - If they underestimate, they charge too little (and lose money)

HOW IS THIS USEFUL?

- They use linear regression
- Have data on how much damage other customers have caused
- Regression analysis of damages caused (Y), depending on age (X)
- Based on your age, predict how much damage you will cause
  - Damages = a + b*age
- That determines your rate

HOW IS THIS USEFUL?

- Linear regression also important for public policy



- Infant mortality rates (Death under 1 year of age per 1,000 live births)

HOW IS THIS USEFUL?

- Some of these rates are appalling
  - Mali: Out of 1,000 babies born alive, 100 die before their first birthday
- If we want to lower infant mortality rates, we need to know what causes them

HOW IS THIS USEFUL?

- Infant mortality rate = 39.9 - 0.0008889*GDP per capita

HOW IS THIS USEFUL?

- Infant mortality rate = 39.9 - 0.0008889*GDP per capita
- For each dollar that GDPpc is higher, infant mortality expected to decrease by 0.0008889
- If GDPpc=0, infant mortality is expected to be 39.9

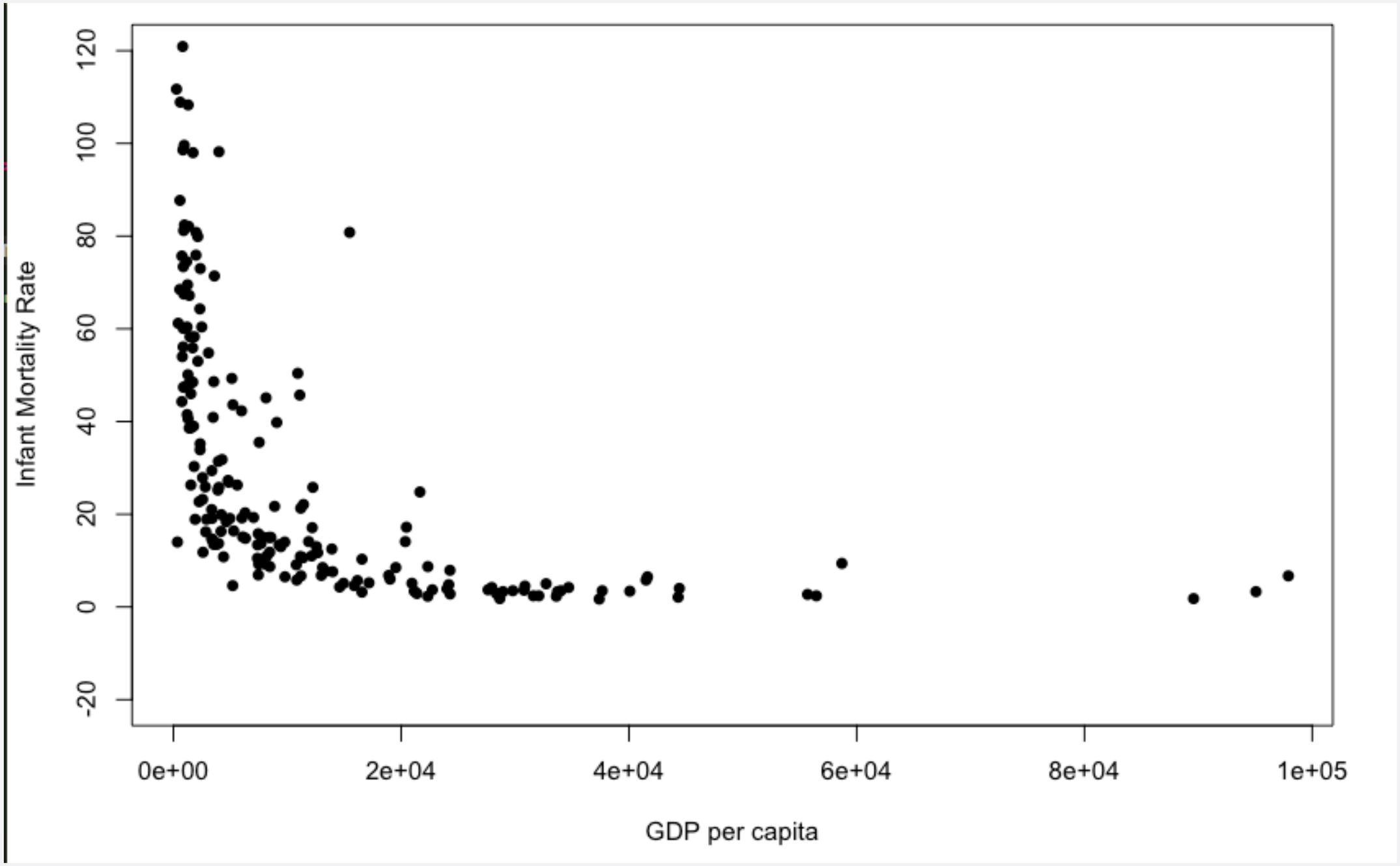
HOW IS THIS USEFUL?

- Infant mortality rate = 39.9 - 0.0008889*GDP per capita
- GDP per capita of Mexico is $10,046
  - Expected rate: 39.9 - 0.0008889*10,046=30.97
- GDP per capita of Mali is $874
  - Expected rate: 39.9 - 0.0008889*874=39.12

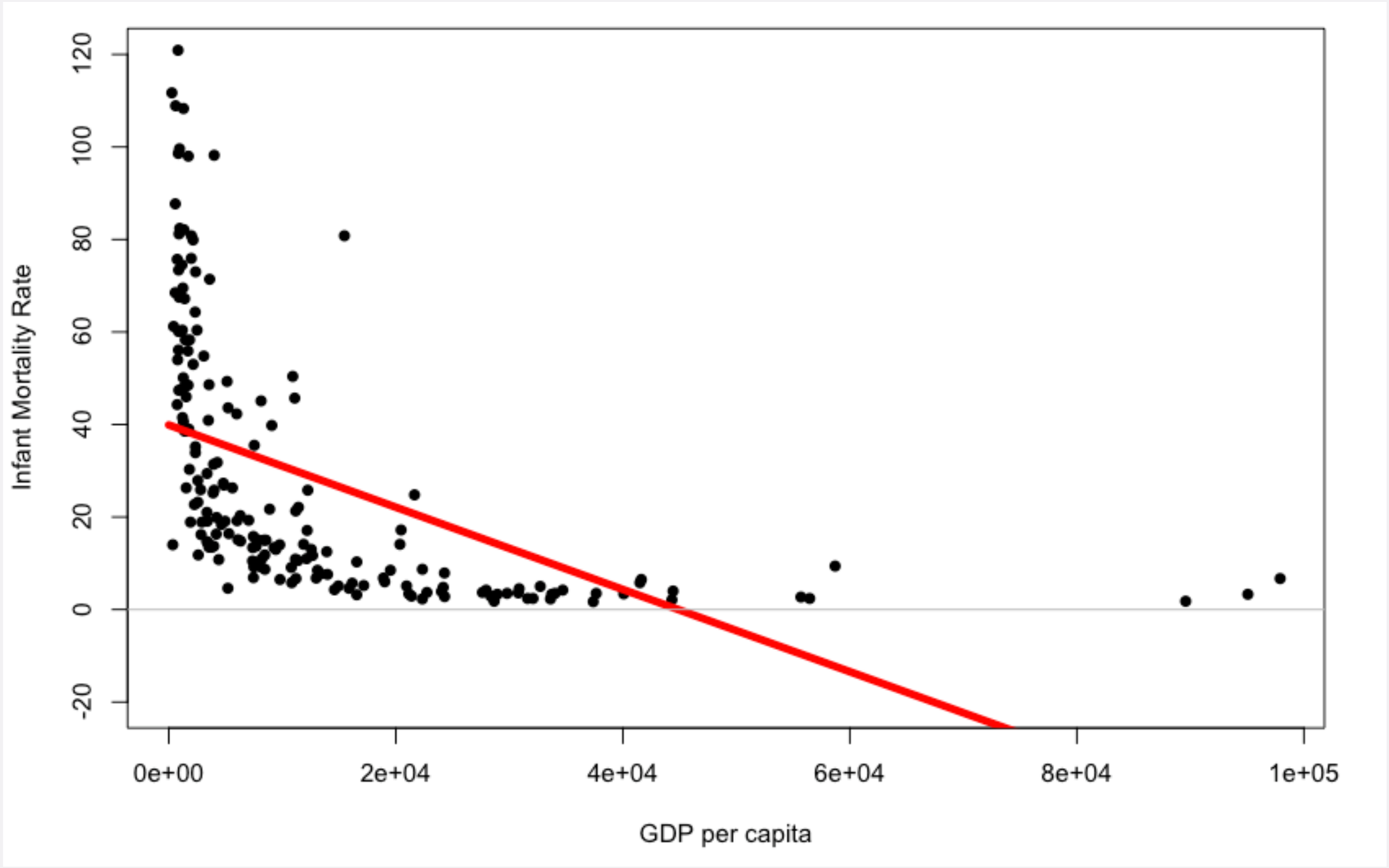
TODAY

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

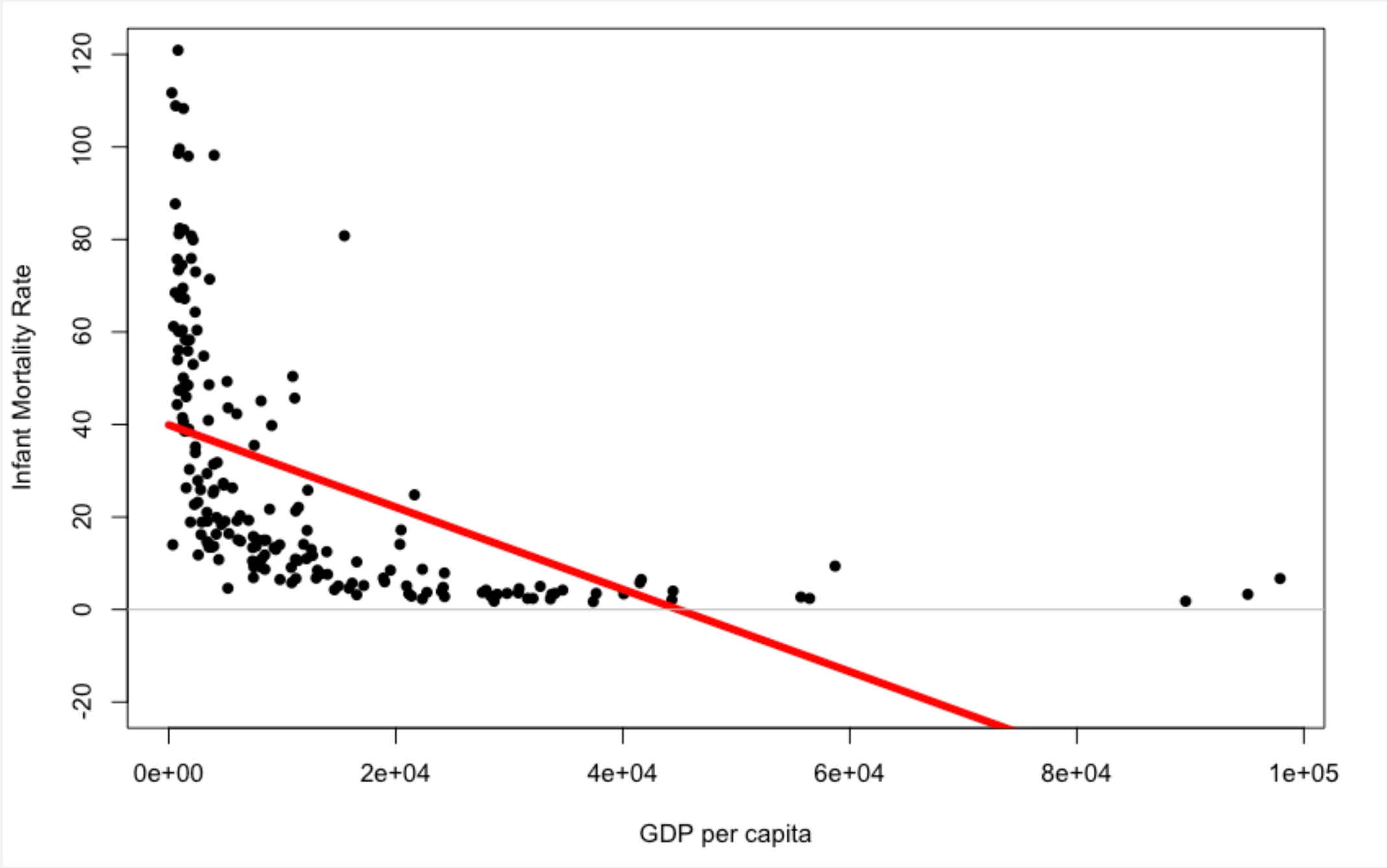
PLOT


Infant Mortality Rate
120
100
80
60
40
20
0
-20
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
GDP per capita

OH NO...

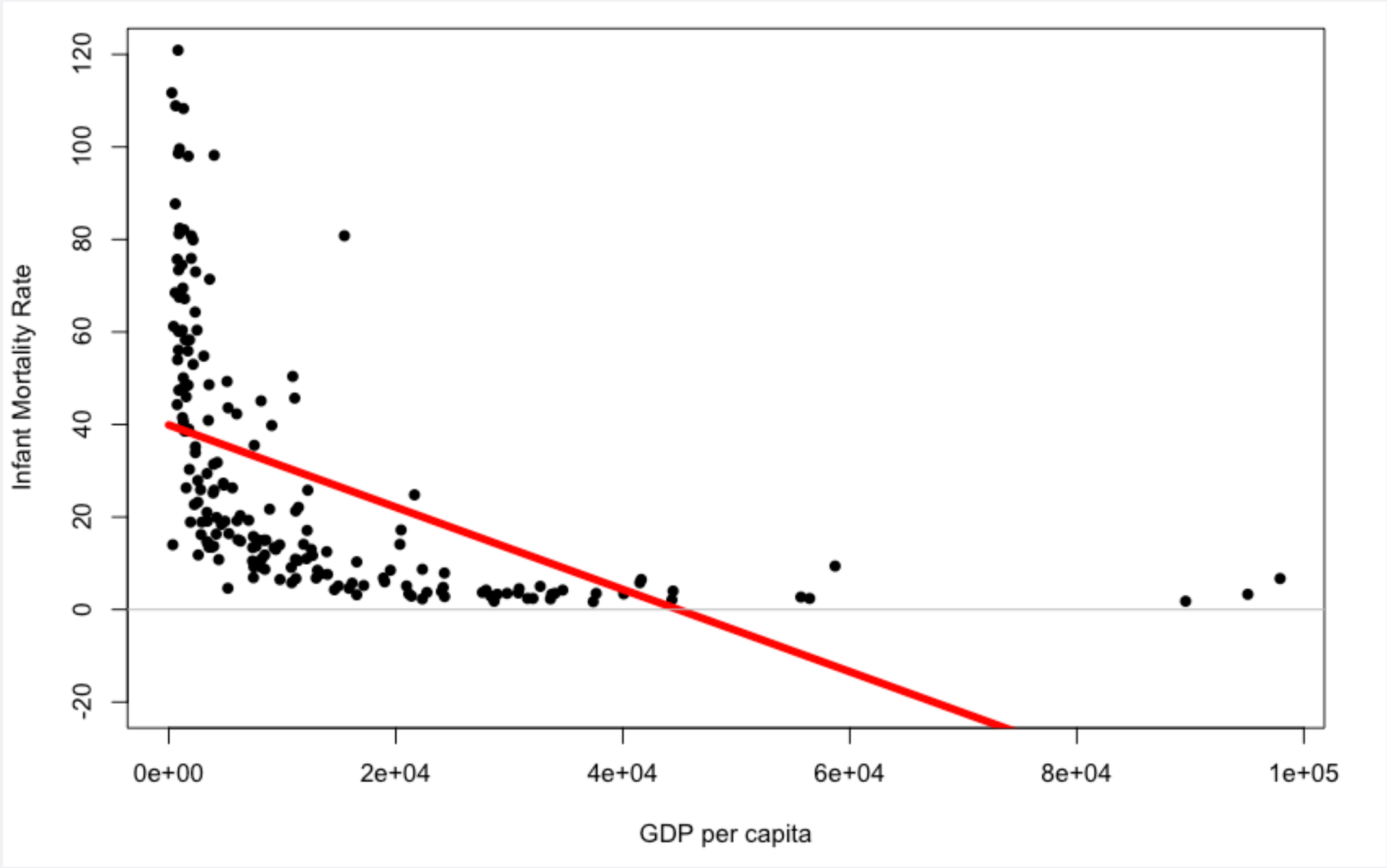

Infant Mortality Rate
120
100
80
60
40
20
0
-20
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
GDP per capita

LINEARITY



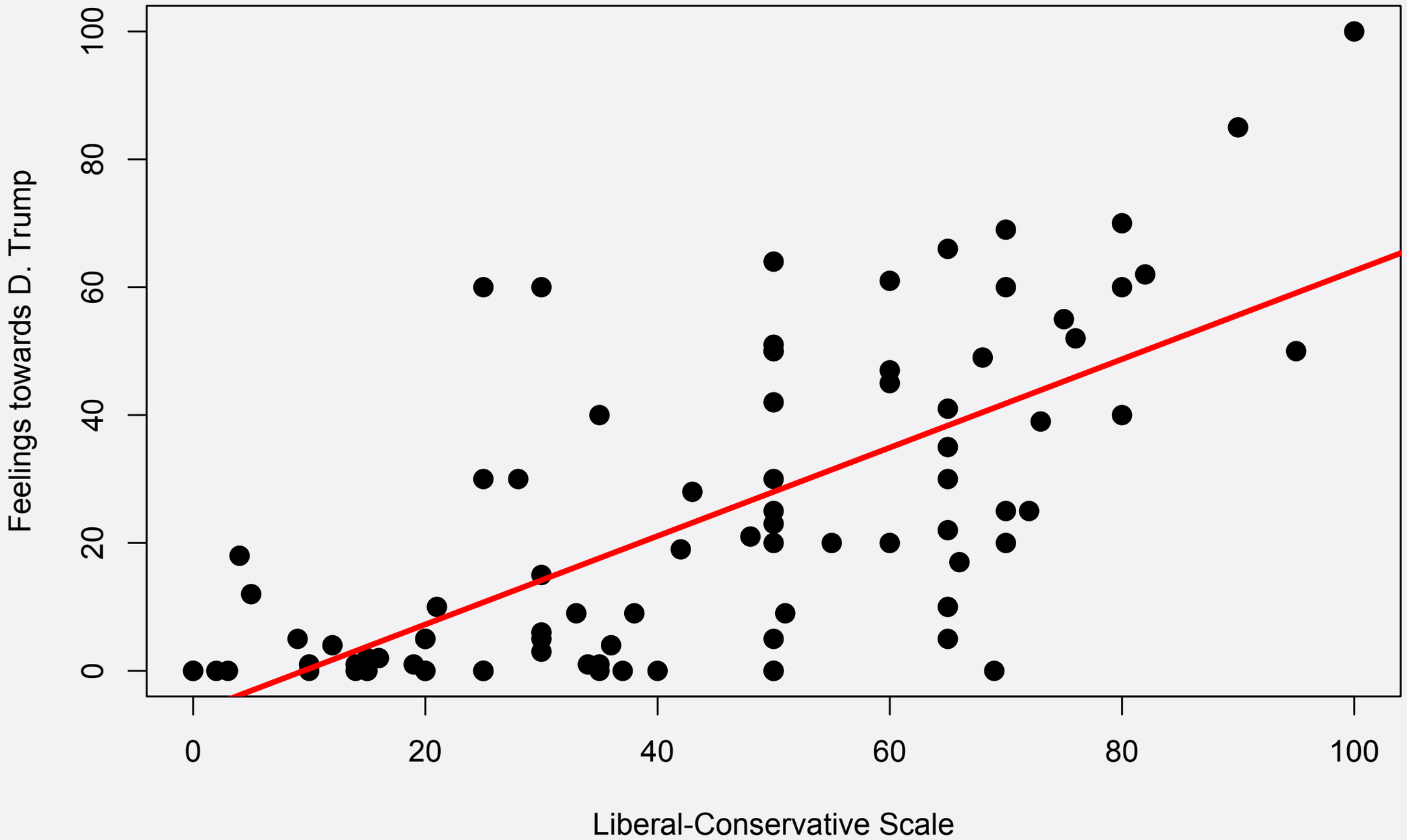
- Linear regression really means *linear*
- Often, effect of x on y is *not* linear

EXPLORING DATA



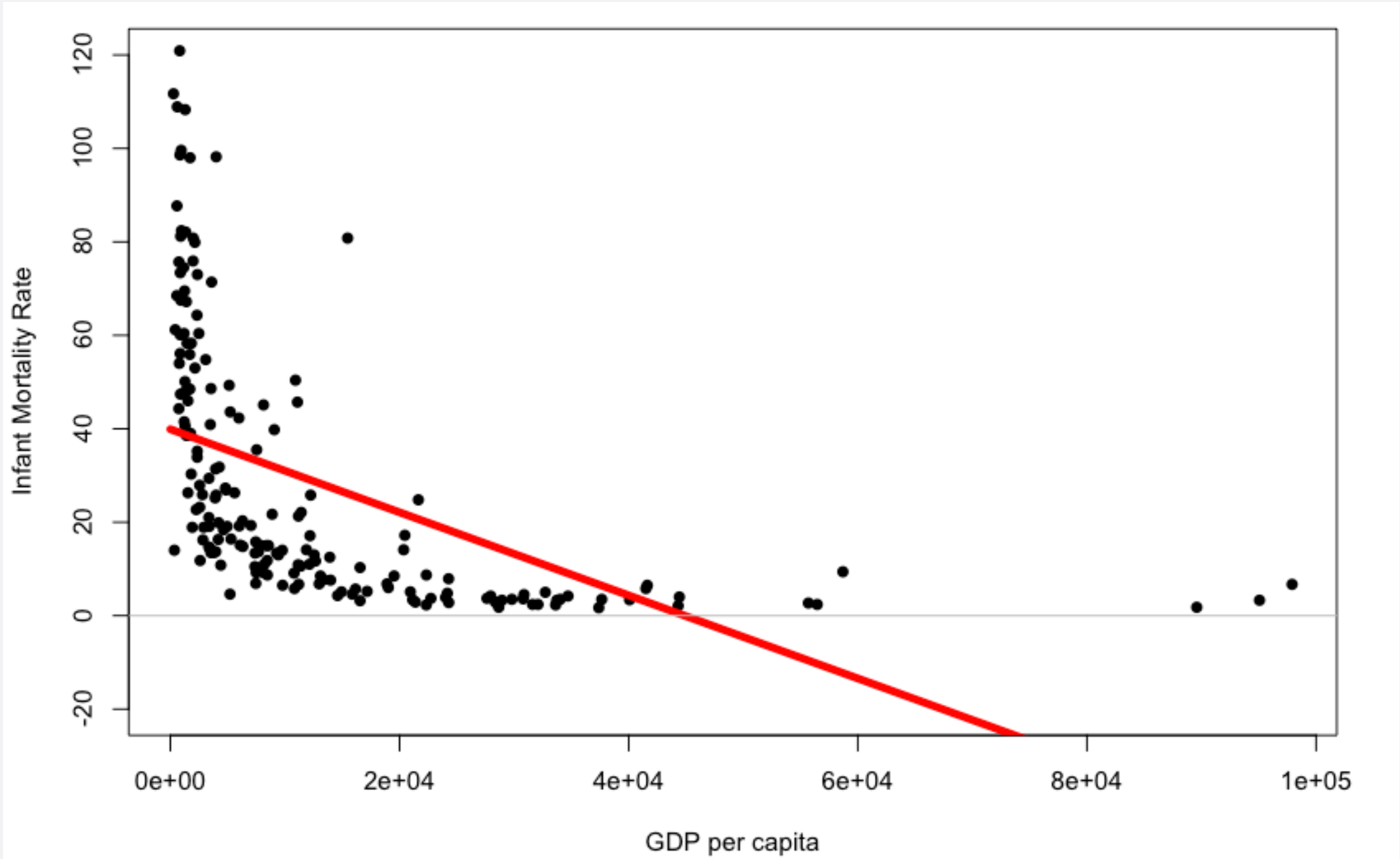
- **Always start an analysis by getting to know your data, make plots etc.**

FROM OUR SURVEY



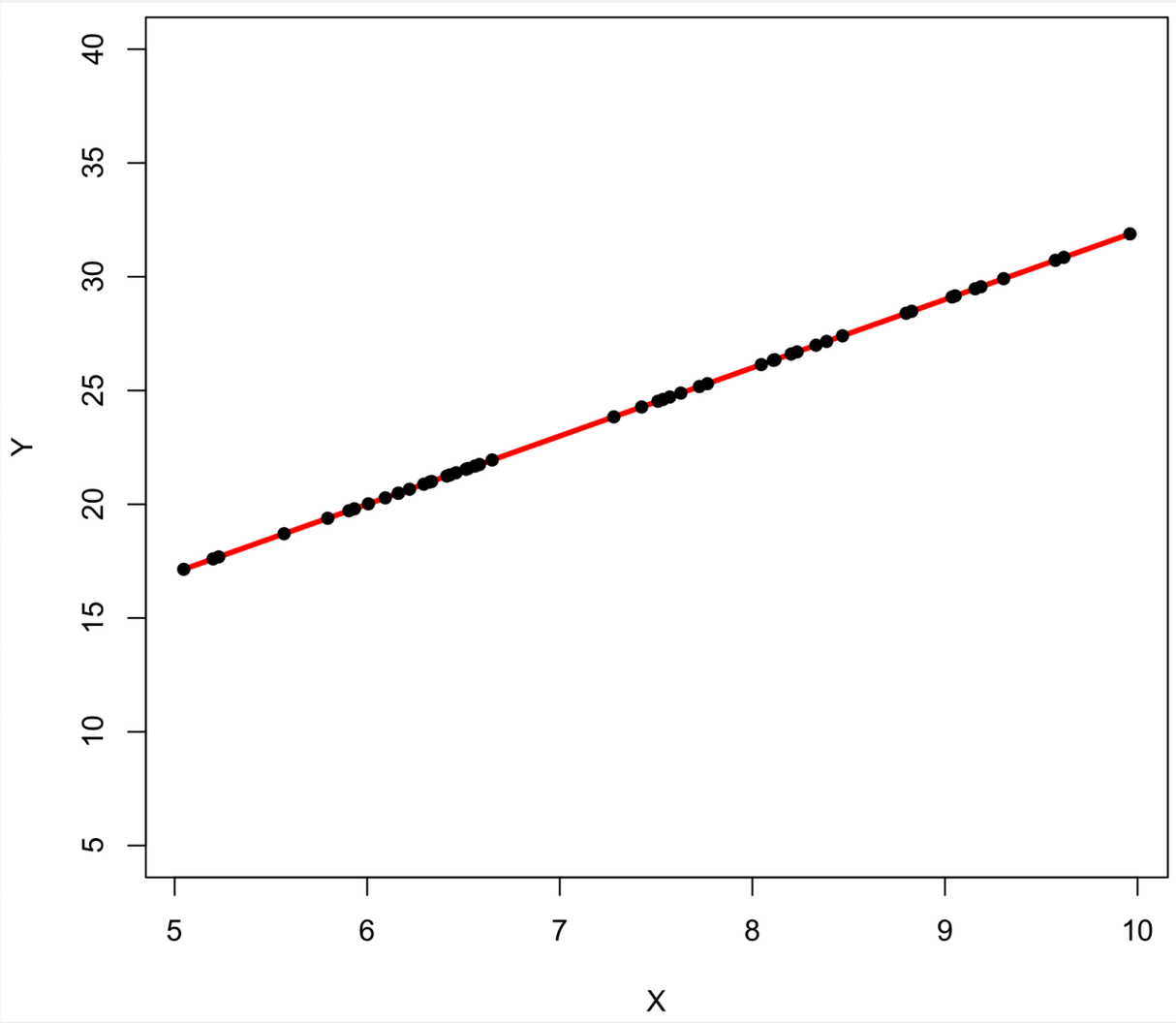
- Score=-6.57 + 0.7 * Lib/Cons
- Intercept is negative!

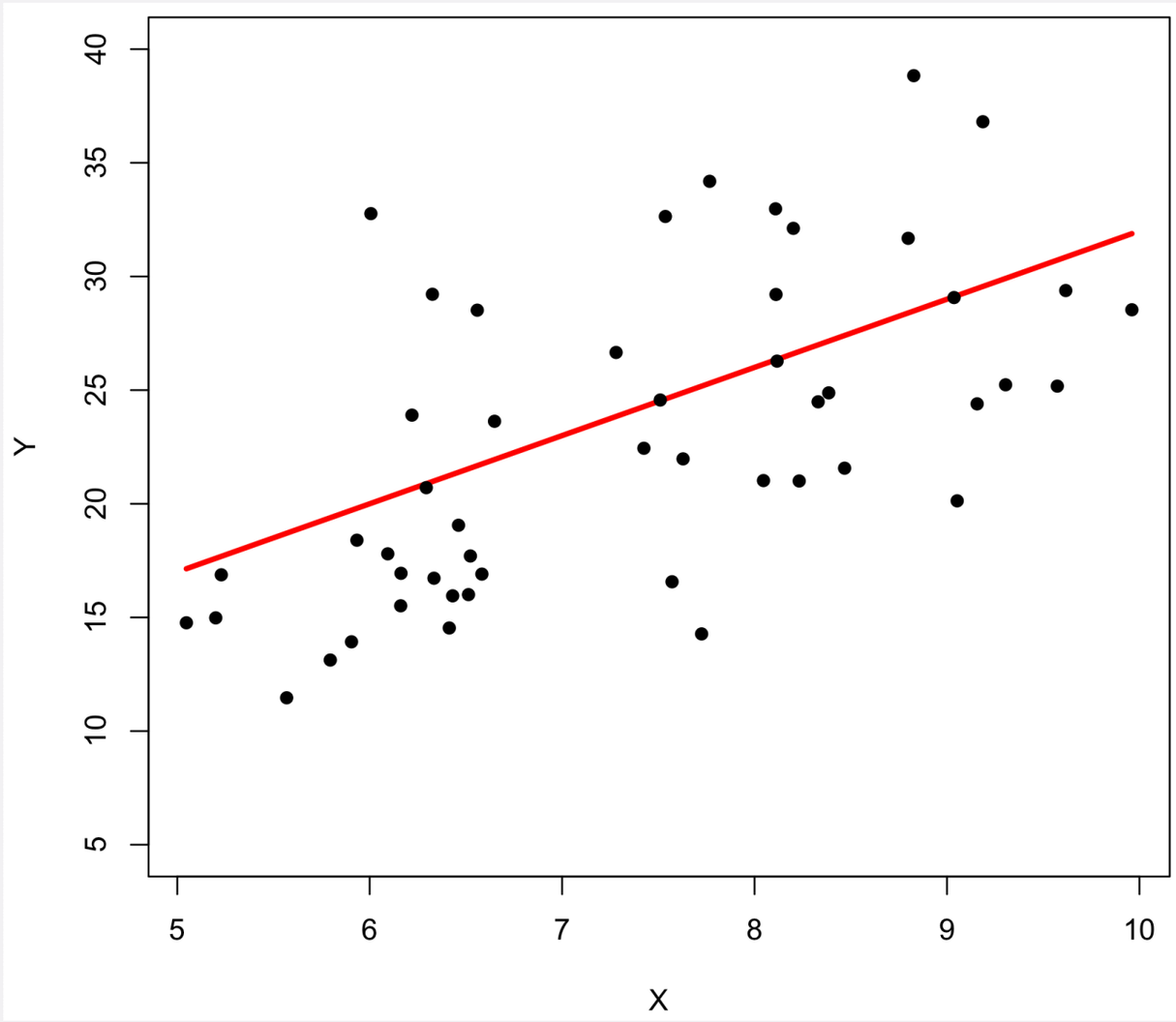
ANOTHER THING



- This line is the line that minimizes squared prediction error
- But: Even this line has a lot of prediction error!

MORE GENERALLY



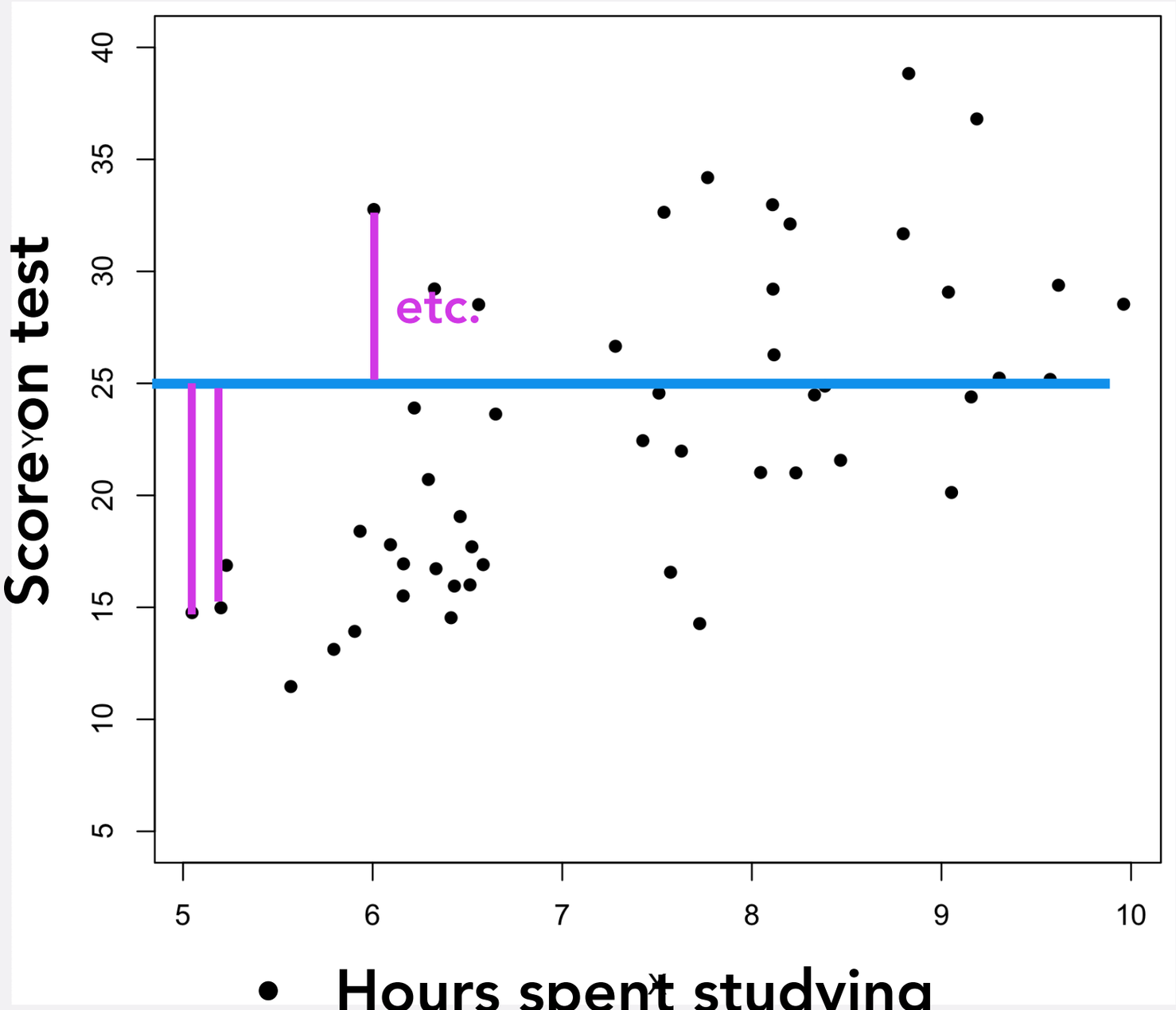


- Same regression equation in both situations
  - Y = 2 + 3*X
- But: X explains Y much better in the first than in the second
- Regression equation does not tell us *how much* it explains

EXPLANATORY POWER MEASURE

- Need: measure of how well independent variable explains dependent variable in a linear regression
- Idea: How much of the variation in Y can we predict using X?

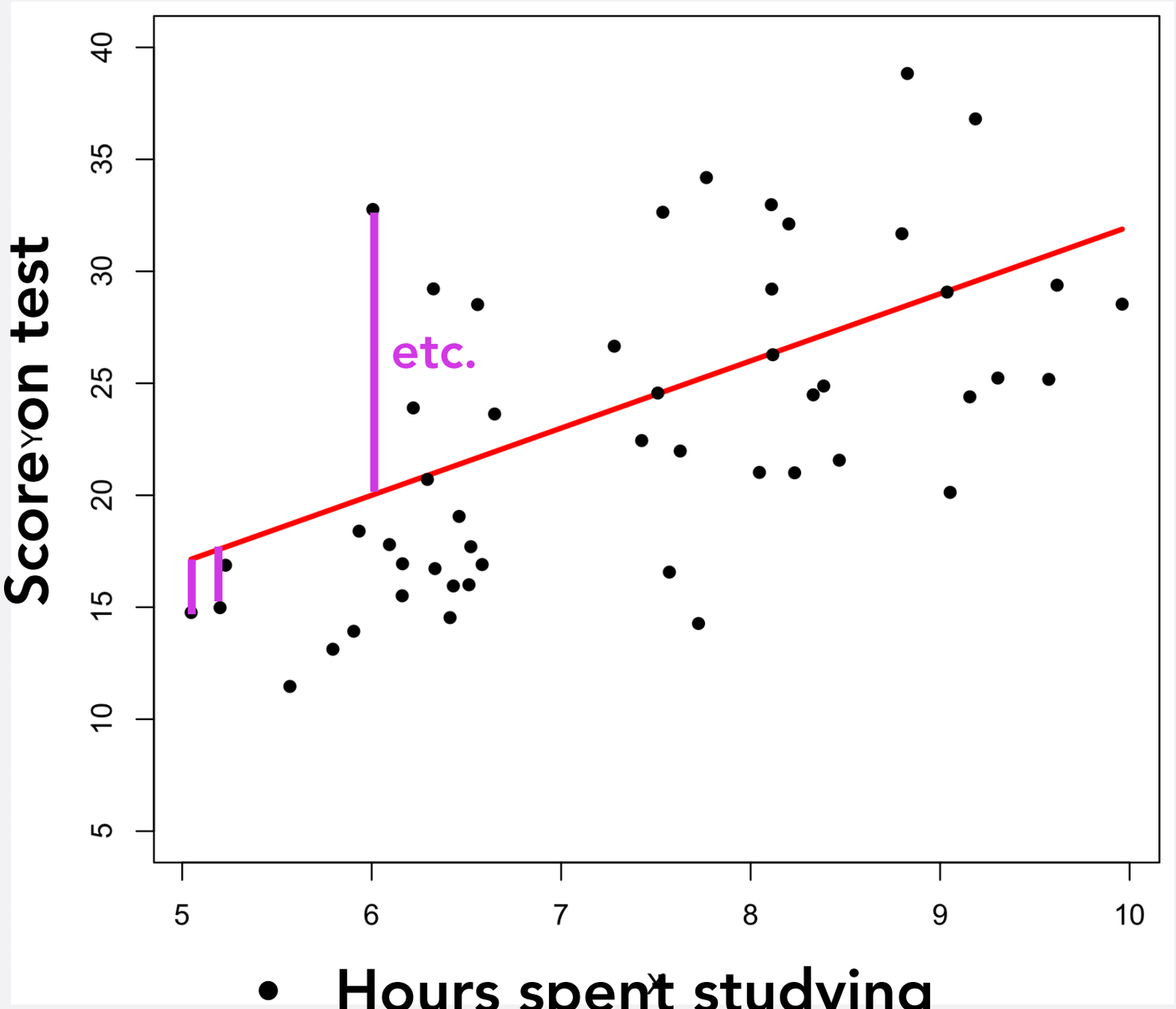
EXPLANATORY POWER MEASURE



- Hours spent studying

- We take mean (25) and compute squared prediction error for each observation
- =Variance of Y (test score): 47.5

EXPLANATORY POWER MEASURE



- Hours spent studying

- Now: We take regression line and compute squared prediction error for each observation
- ="Residual variance"=29.6

EXPLANATORY POWER MEASURE

- Squared prediction error without regression line: 47.5
- Squared prediction error with regression line (for hours spent studying): 29.6
- 29.6 is 62.3% of 47.5
  - So we were able to reduce squared prediction error by 100-62.3=37.7%
  - In other words, hours spent studying explains 37.7% of variance in test scores

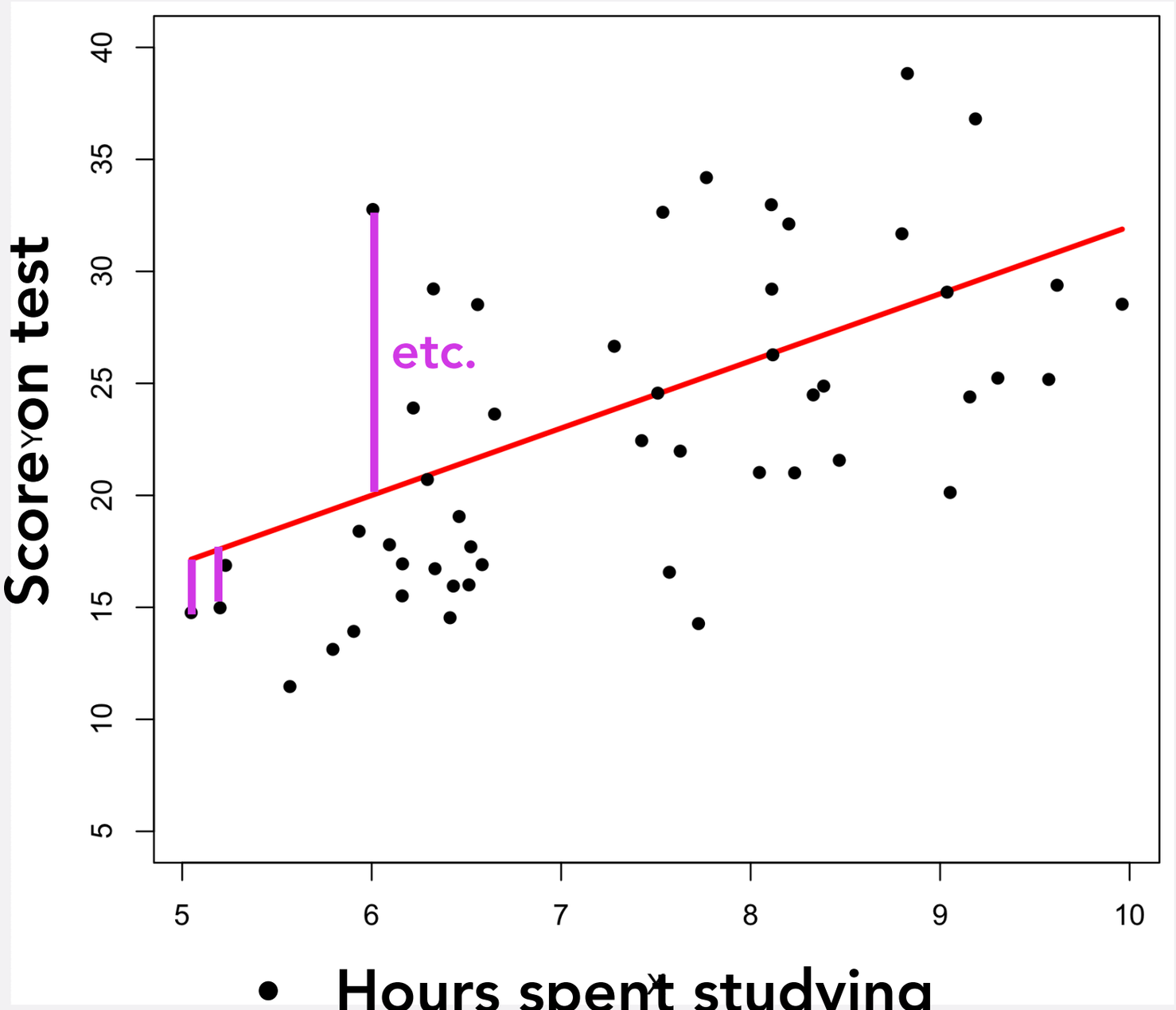
R-SQUARE

- Measure is called R^2
- Interpretation: R^2 tells us how much variation of the dependent variable is explained by the independent variable

R-SQUARE

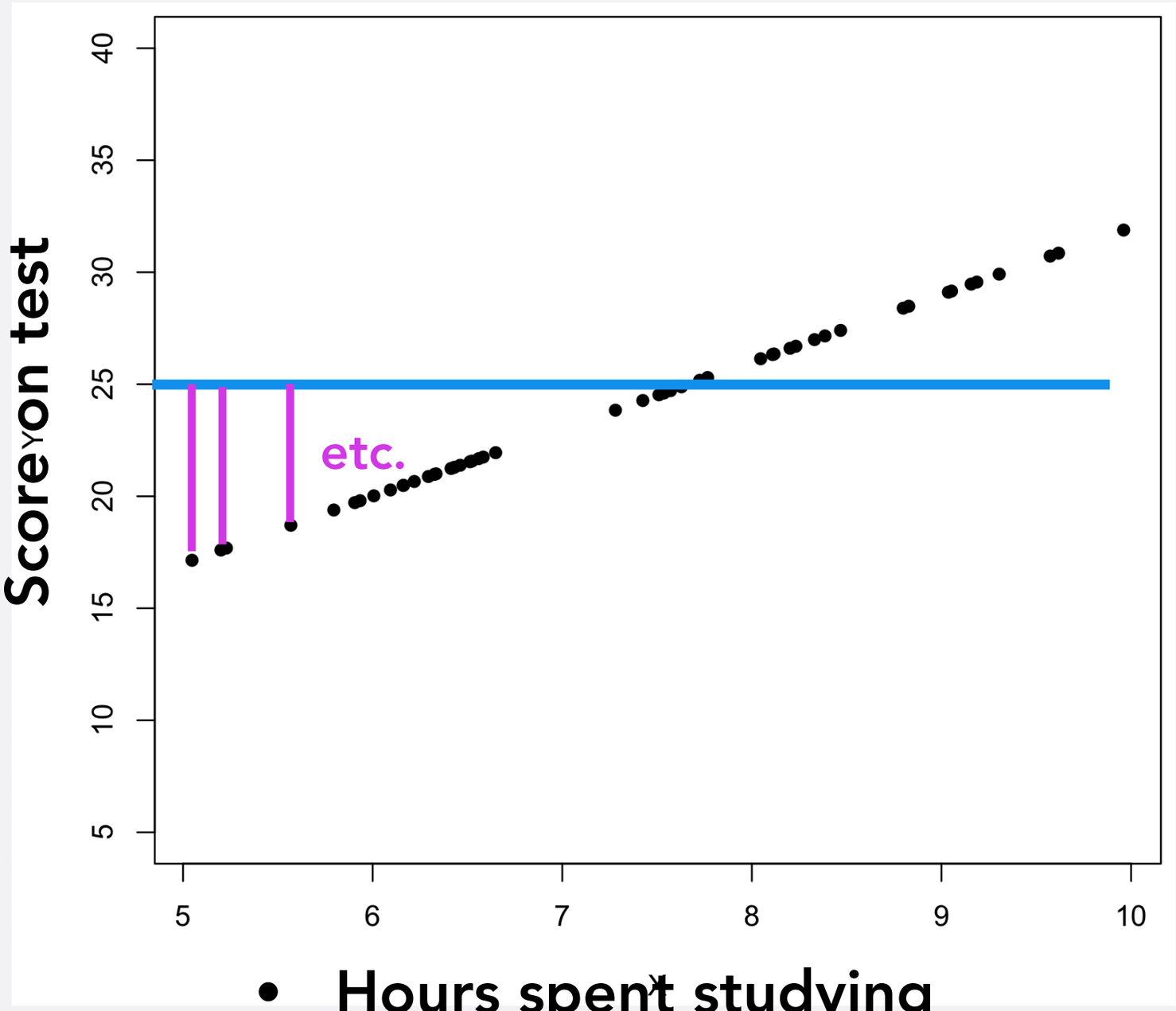
- Typically, not expressed as percentage (between 0 and 100), but as fraction (between 0 and 1)
- 0: The independent variable explains *none* of the variation in the dependent variable
- 1: The independent variable explains *all* of the variation in the dependent variable

EXPLANATORY POWER MEASURE



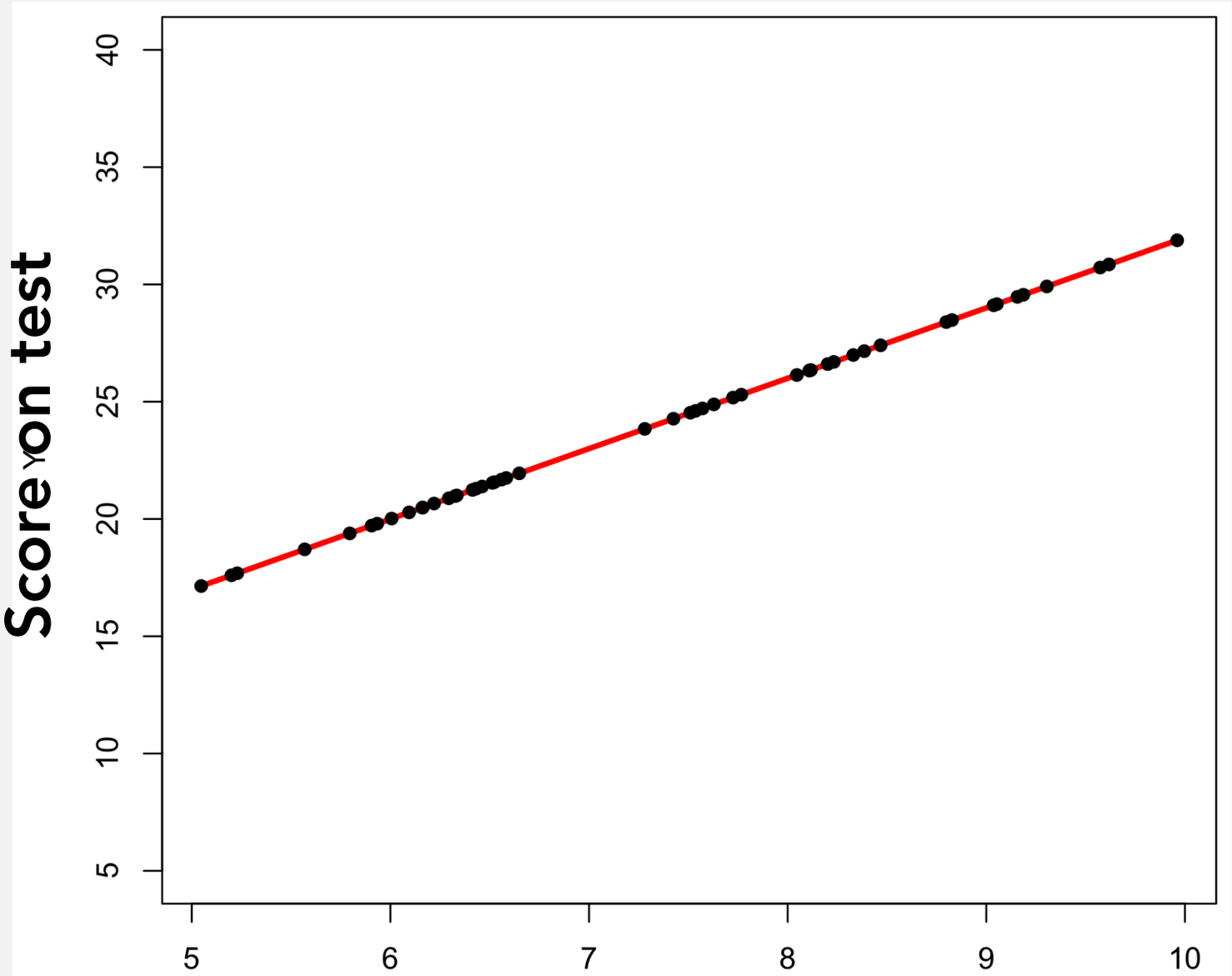
- Hours spent studying explains 37.7% of variance in test scores
- So: $R^2 = 0.377$

EXPLANATORY POWER MEASURE



- Variance of Y (test score): 15.7

EXPLANATORY POWER MEASURE



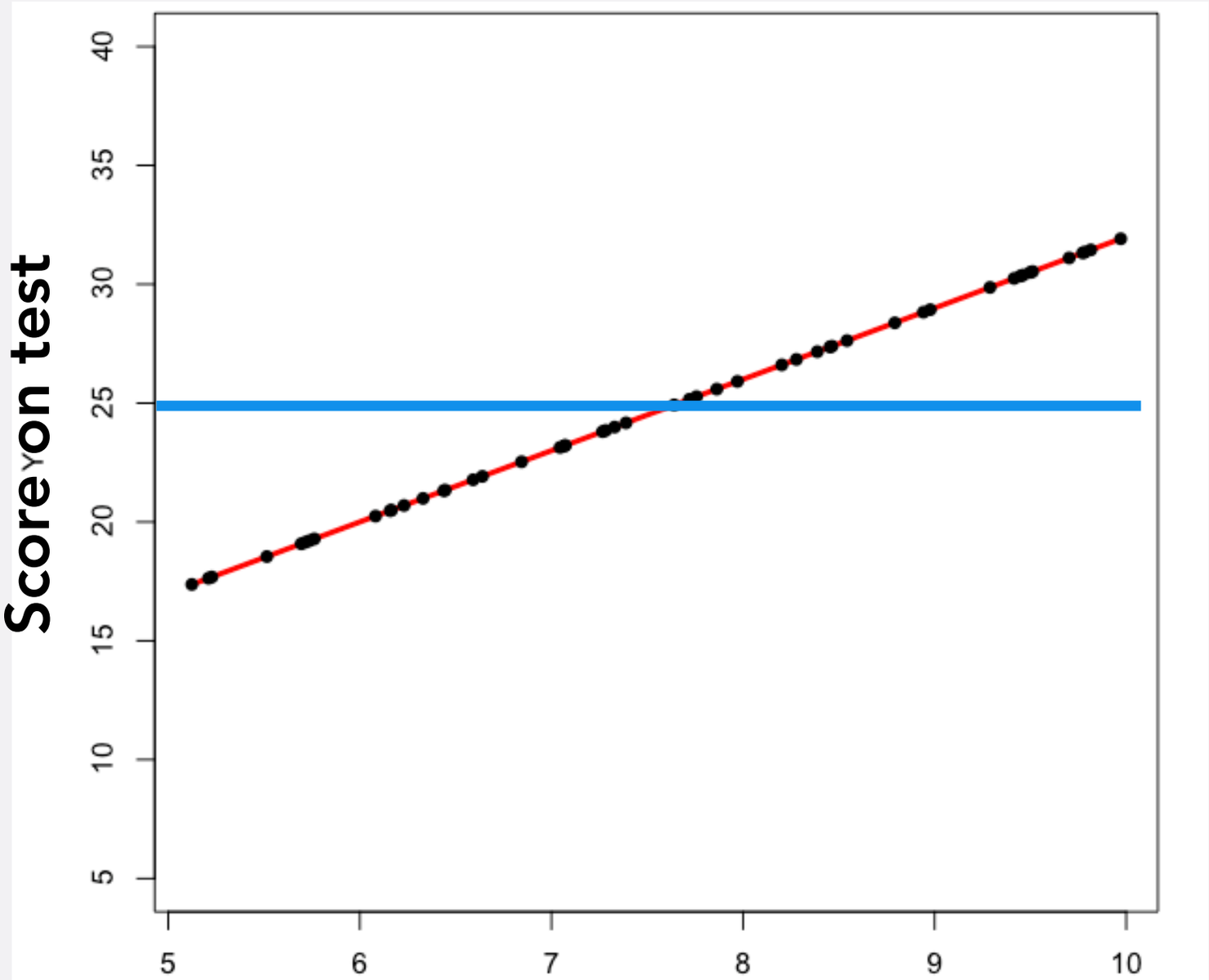
- Hours spent studying

- Residual variance: 0

EXPLANATORY POWER MEASURE

- Squared prediction error without regression line: 15.7
- Squared prediction error with regression line (for hours spent studying): 0
- So we were able to reduce squared prediction error by 100%
- Hours spent studying explains 100% of variance in test scores
- $R^2=1$

EXPLANATORY POWER MEASURE



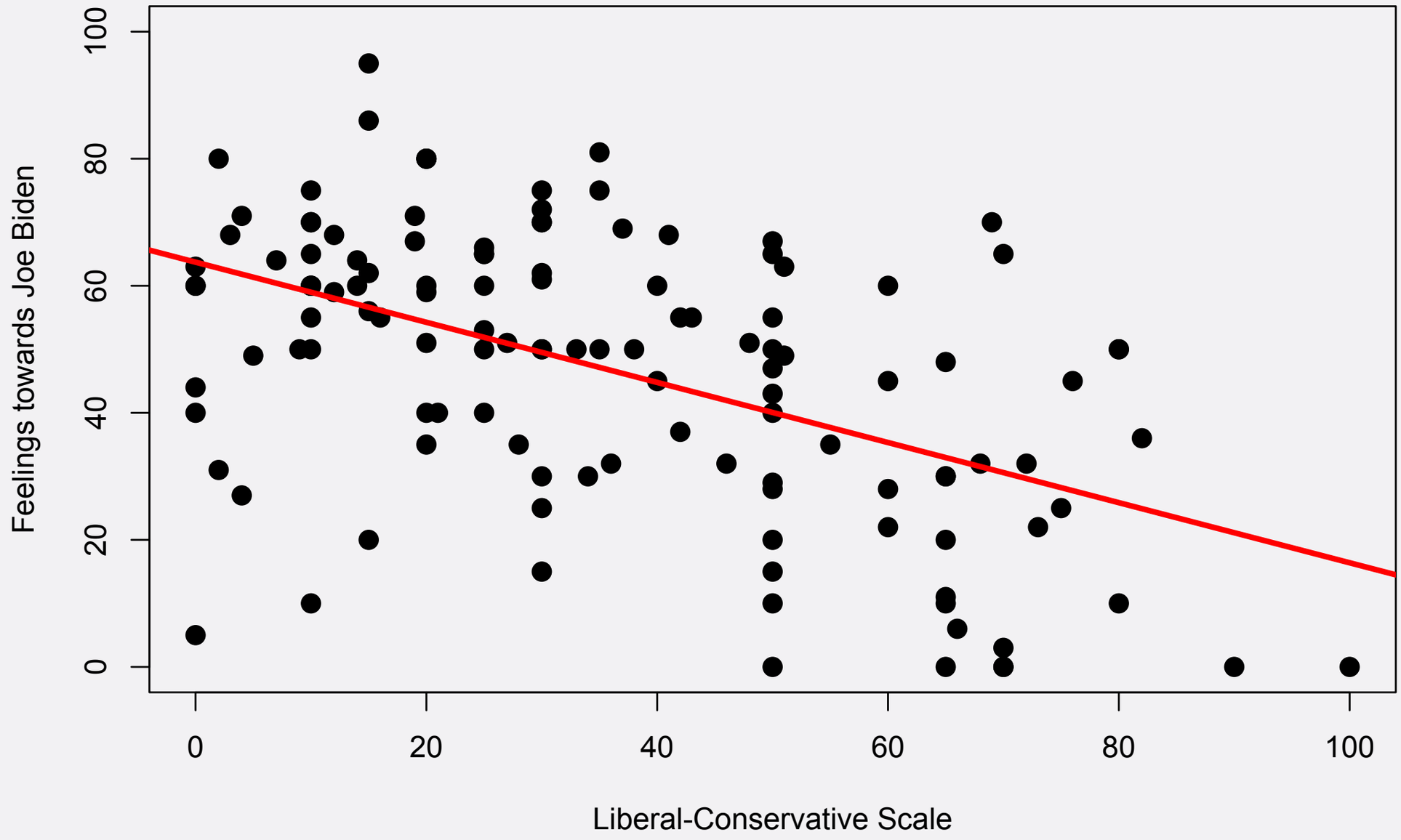
- Hours spent studying explains *all* variation in scores

BACK TO OUR EXAMPLE

```
> summary(lm(therm_2 ~ libcons_1, data = data))

Call:
lm(formula = therm_2 ~ libcons_1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-58.713 -12.954   1.019  12.484  38.939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.71327    3.09127  20.611  < 2e-16 ***
libcons_1   -0.47323    0.07174  -6.597 1.12e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.23 on 123 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.2613,	Adjusted R-squared:  0.2553
F-statistic: 43.51 on 1 and 123 DF,  p-value: 1.117e-09
```

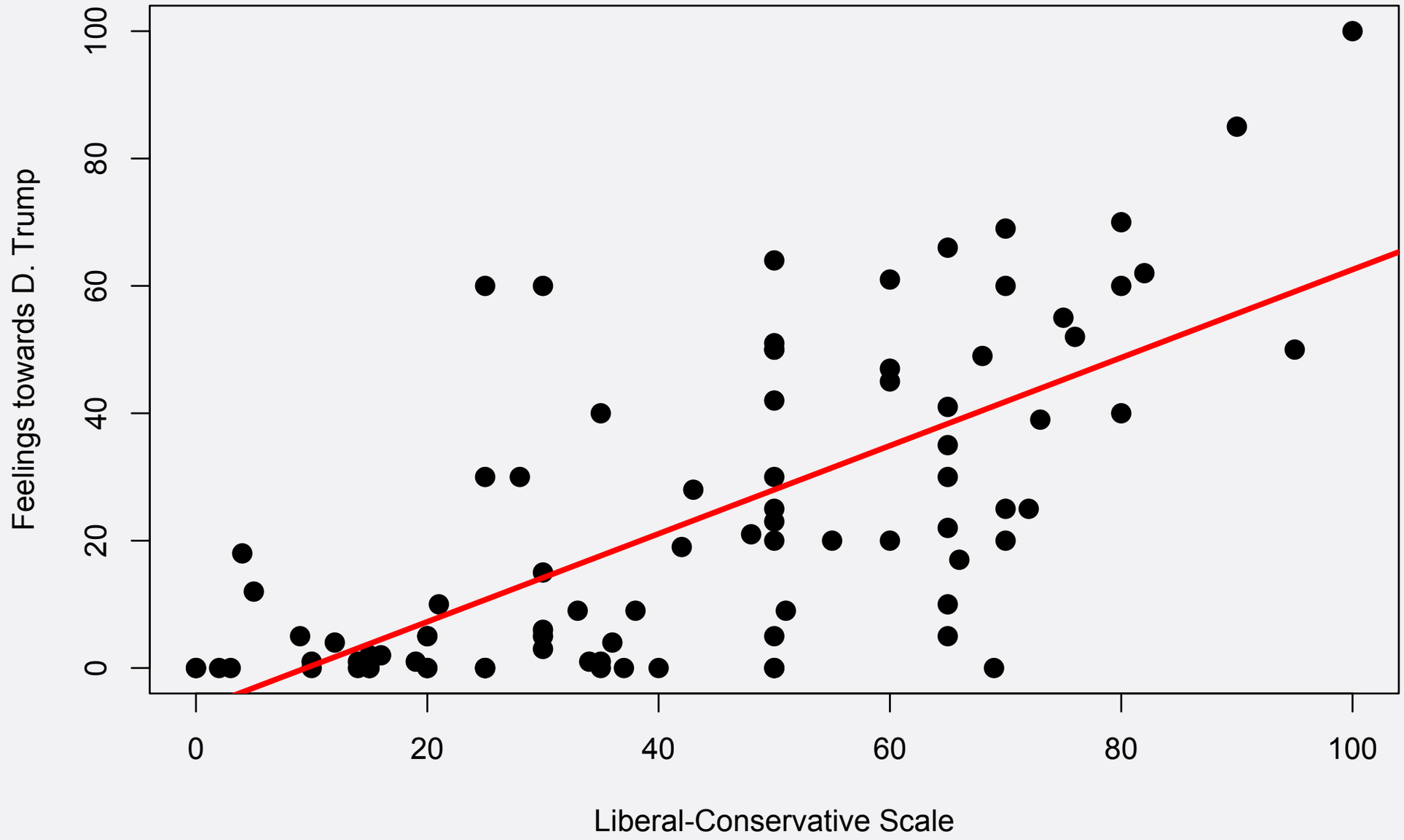
- DV: Rating of J. Biden (therm_2)
- IV: Liberal-conservative scale (libcons_1)

JOE BIDEN



- Here: R^2=0.26
- Lib/cons rating explains 26% of variance in ratings of J. Biden
- So 74% remain unexplained…

DONALD TRUMP



- Here: R^2=0.48
- Lib/cons rating explains 48% of variance in ratings of D. Trump
- So only 52% remain unexplained…

WHAT WE CAN DO

- We can now...
  - Estimate how much an independent variable X affects a dependent variable Y
  - Tell how much of the variance in Y is explained by X

BIVARIATE RELATIONSHIPS

Dependent Variable	Independent Variable: Nominal/Ordinal	Independent Variable: Interval
Nominal/Ordinal	Cross-Tabulation	Not In This Class…
Interval	Mean Comparison	Correlation Coefficient, Linear Regression

NEXT TIME...

- Is the effect of lib/cons on ratings of J. Biden real?
- Or is it only something that we found in our sample, but lib/cons actually has no effect in the population?