PSC 202
SYRACUSE UNIVERSITY

INTRODUCTION TO POLITICAL ANALYSIS

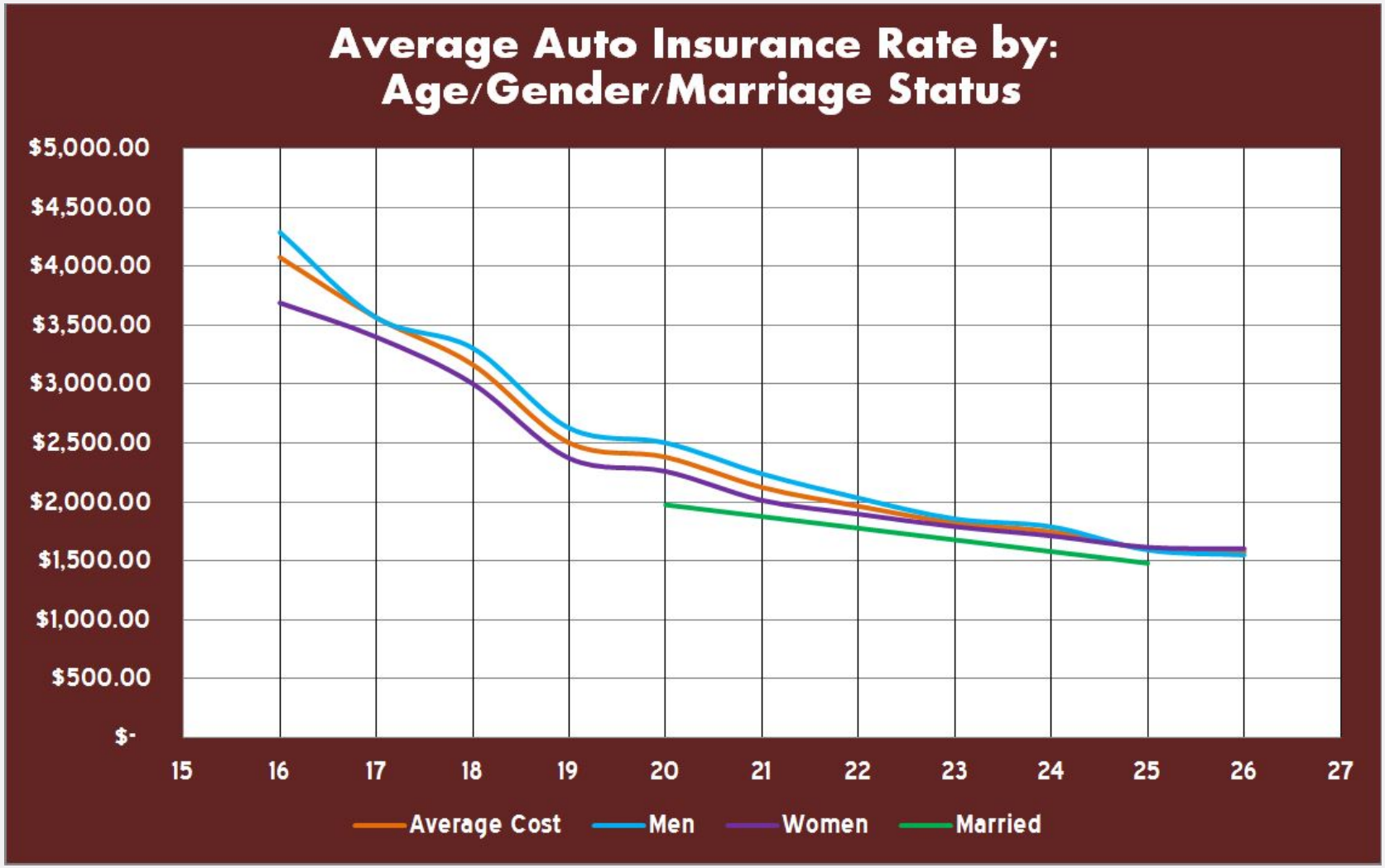
MORE BIVARIATE HYPOTHESIS TESTING, HYPOTHESIS TESTING WITH A SAMPLE

MORE ON REGRESSION LINE

- How is linear regression useful?
- Caveats about linear regression

HOW IS THIS USEFUL?

- Linear regression widely used in private sector



HOW IS THIS USEFUL?

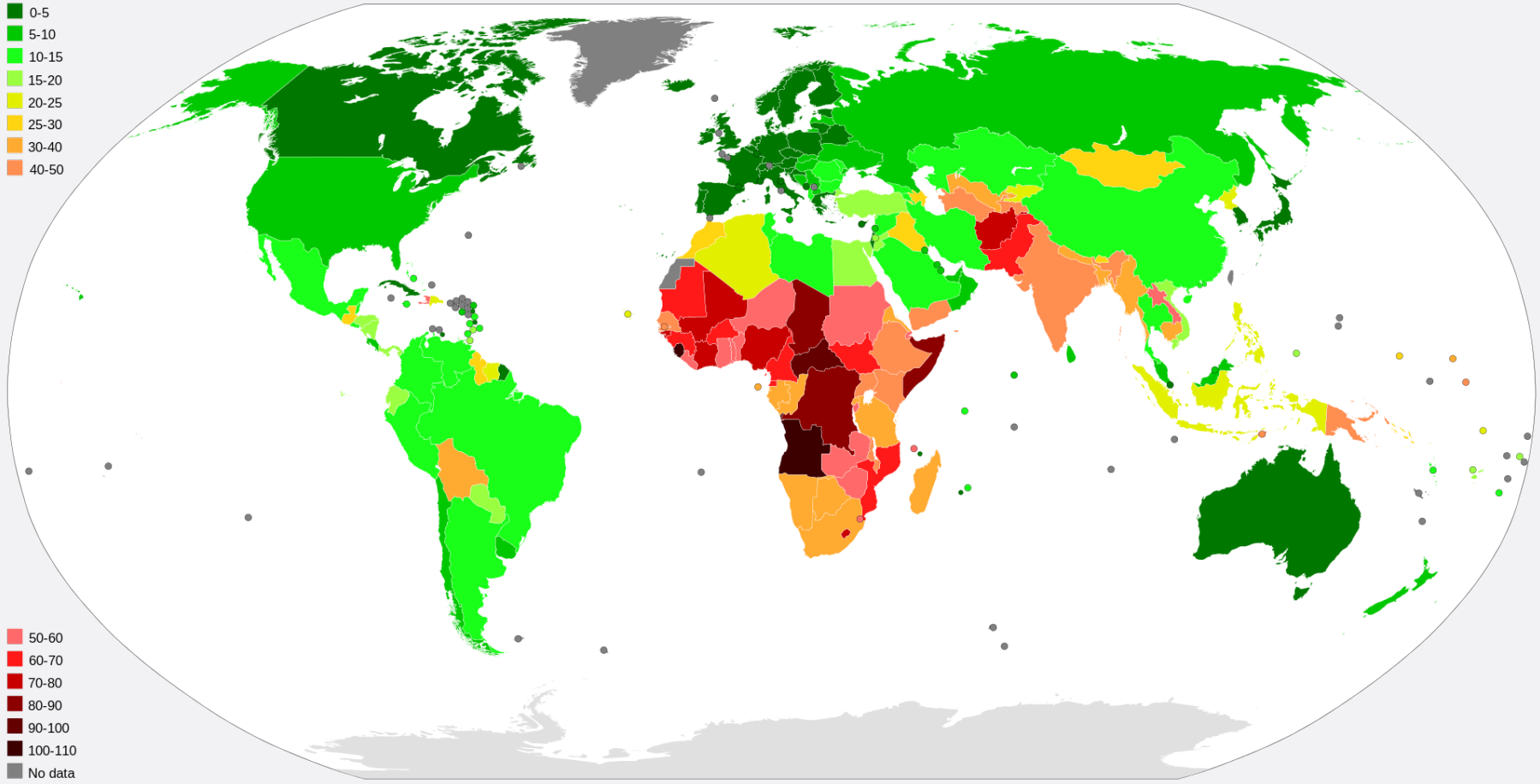
- Insurance company has to decide how much to charge you
- How much to charge you depends on how much in damages they expect to have to pay for you
- Guessing won't do
  - If they overestimate how much damage someone will cause, they charge too much (and the person might buy insurance elsewhere)
  - 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 widely used in public policy research



- 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 the U.S. is $41,627
    - Expected rate: 39.9 - 0.0008889*41,627=2.90

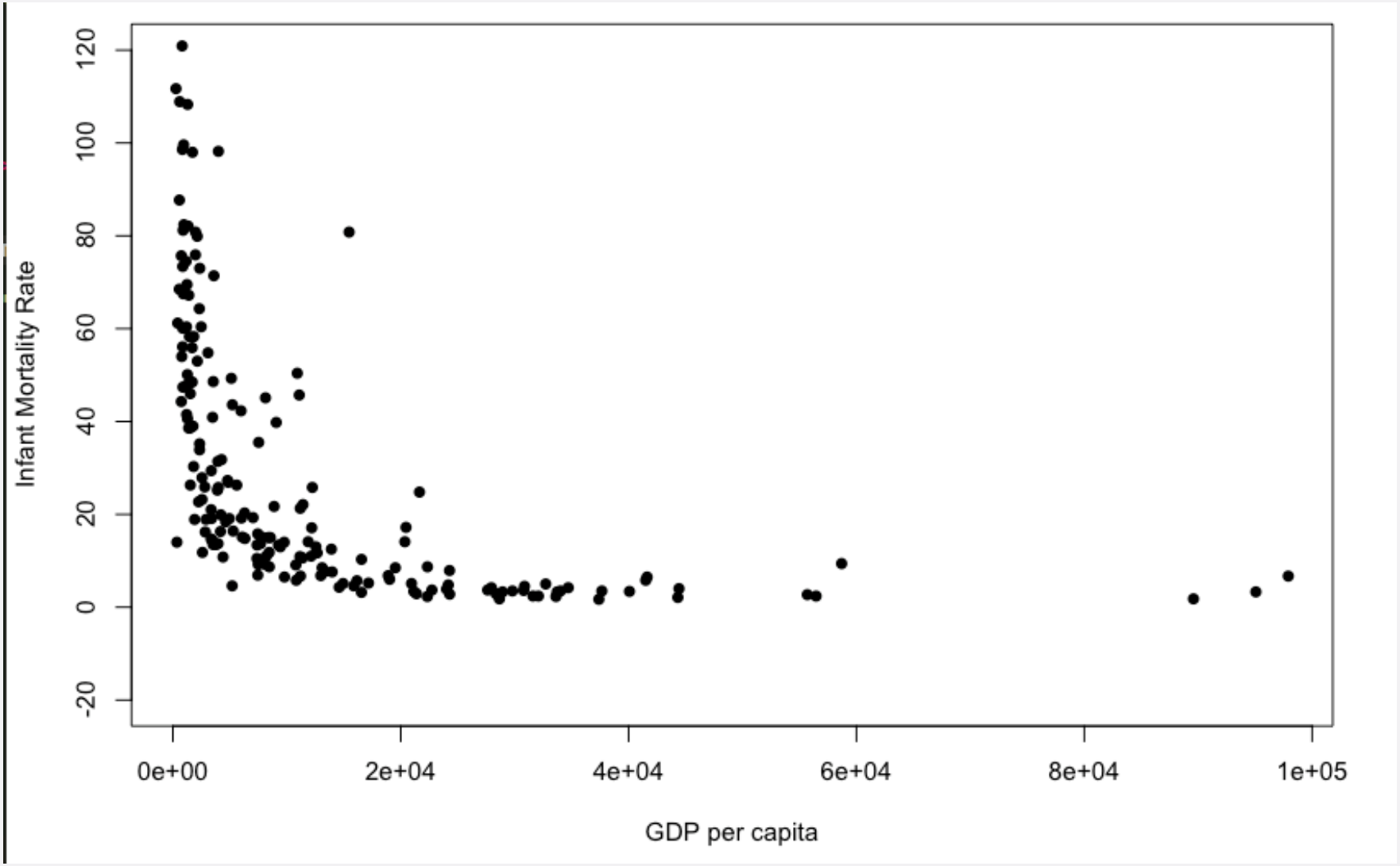
HOW IS THIS USEFUL?

- Infant mortality rate = 39.9 - 0.0008889*GDP per capita
  - GDP per capita of the U.S. is $41,627
    - Expected rate: 39.9 - 0.0008889*41,627=2.90
  - GDP per capita of Mexico is $11,877
    - Expected rate: 39.9 - 0.0008889*11,877=29.34

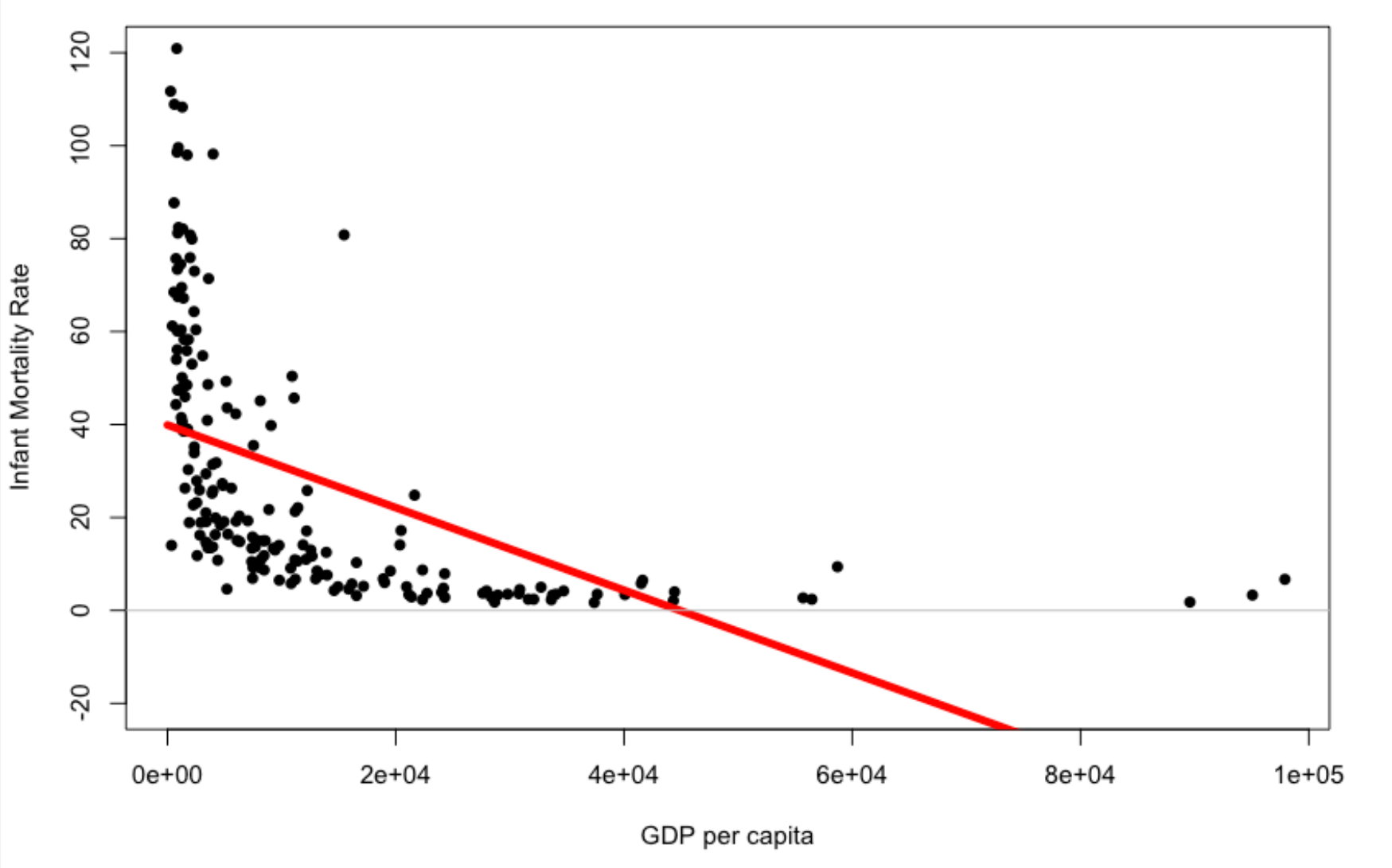
TODAY

- How is linear regression useful?
- Caveats about linear regression

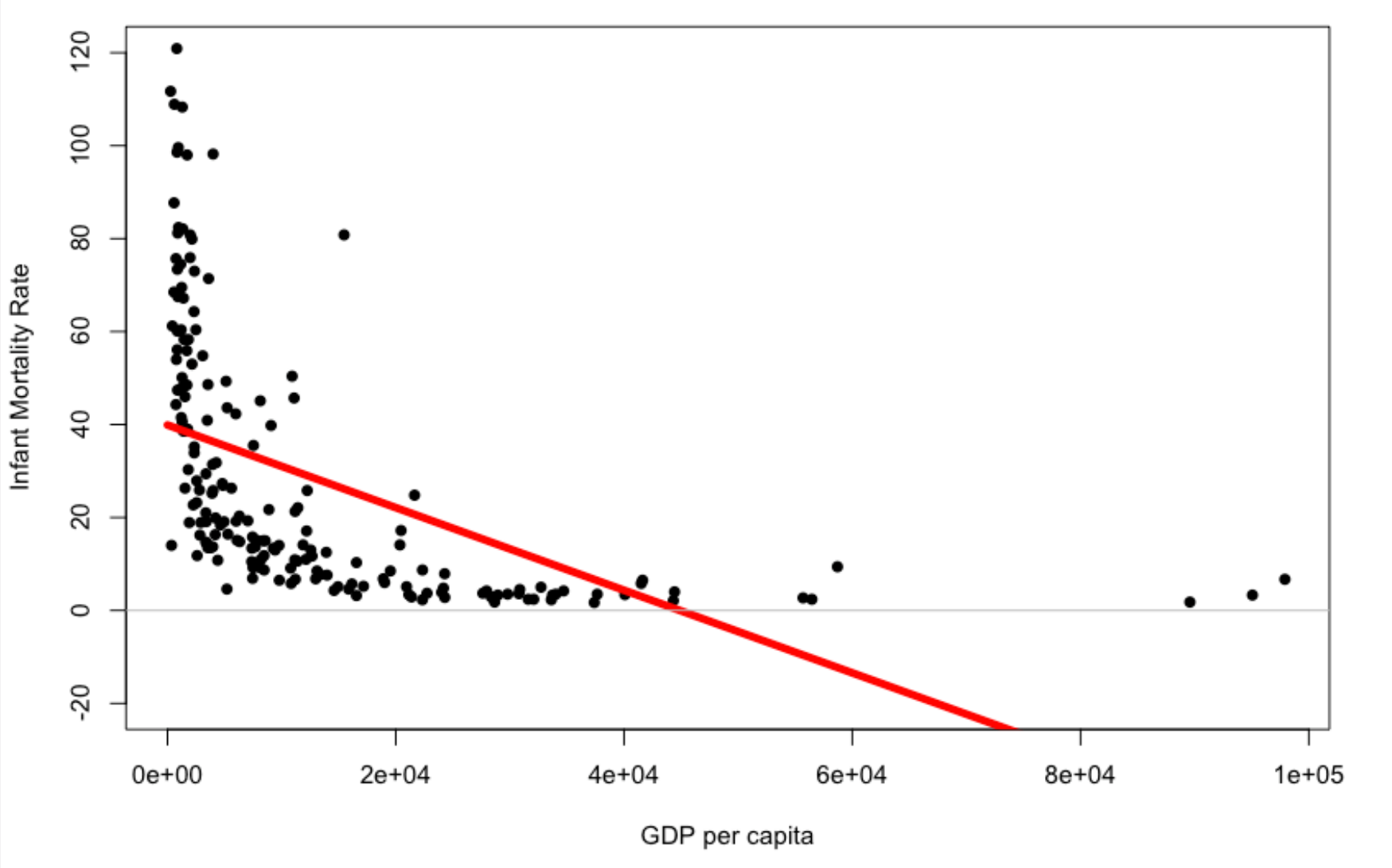
PLOT


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...

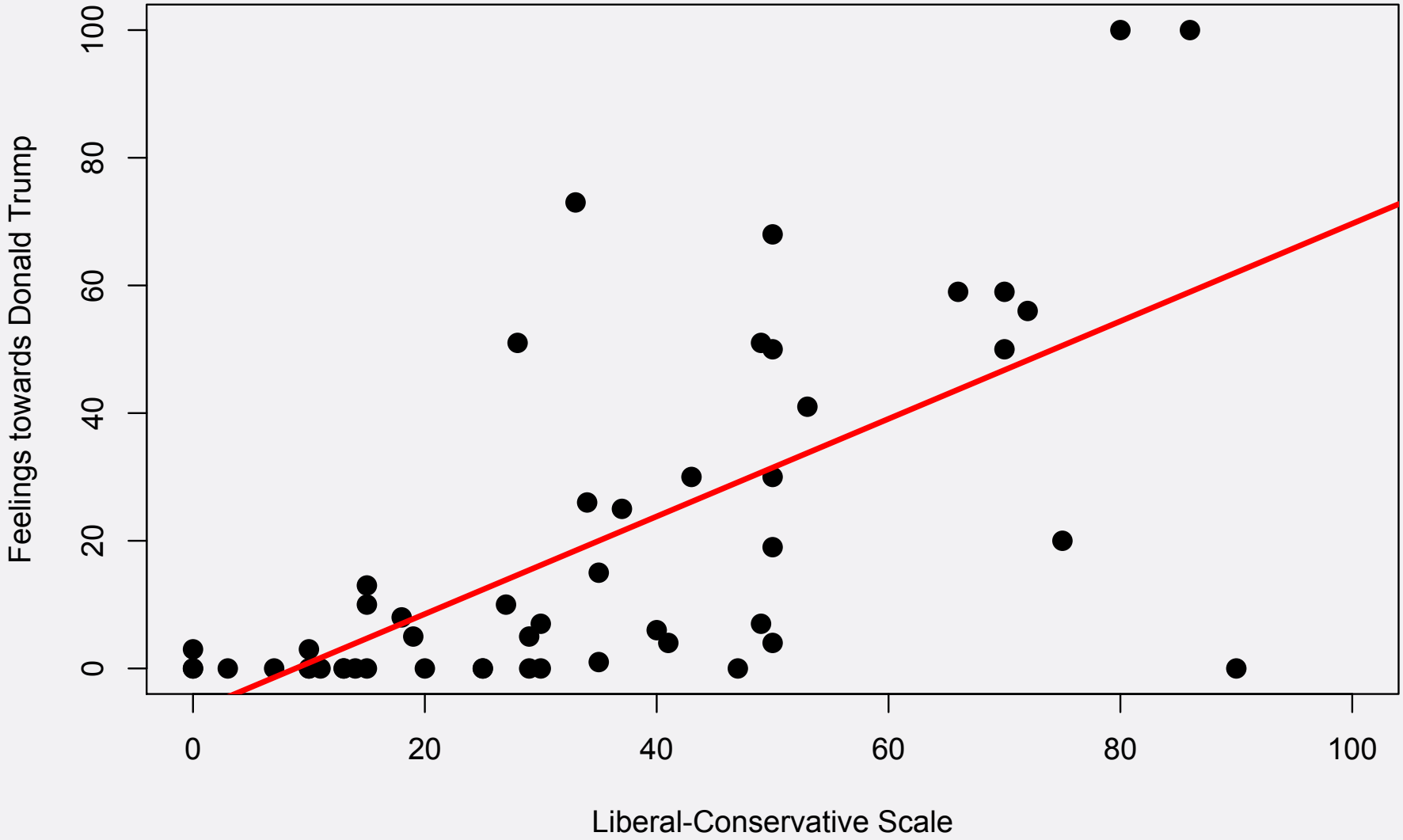

Infant Mortality Rate
120
100
80
60
40
20
0
-20
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
GDP per capita

LINEARITY



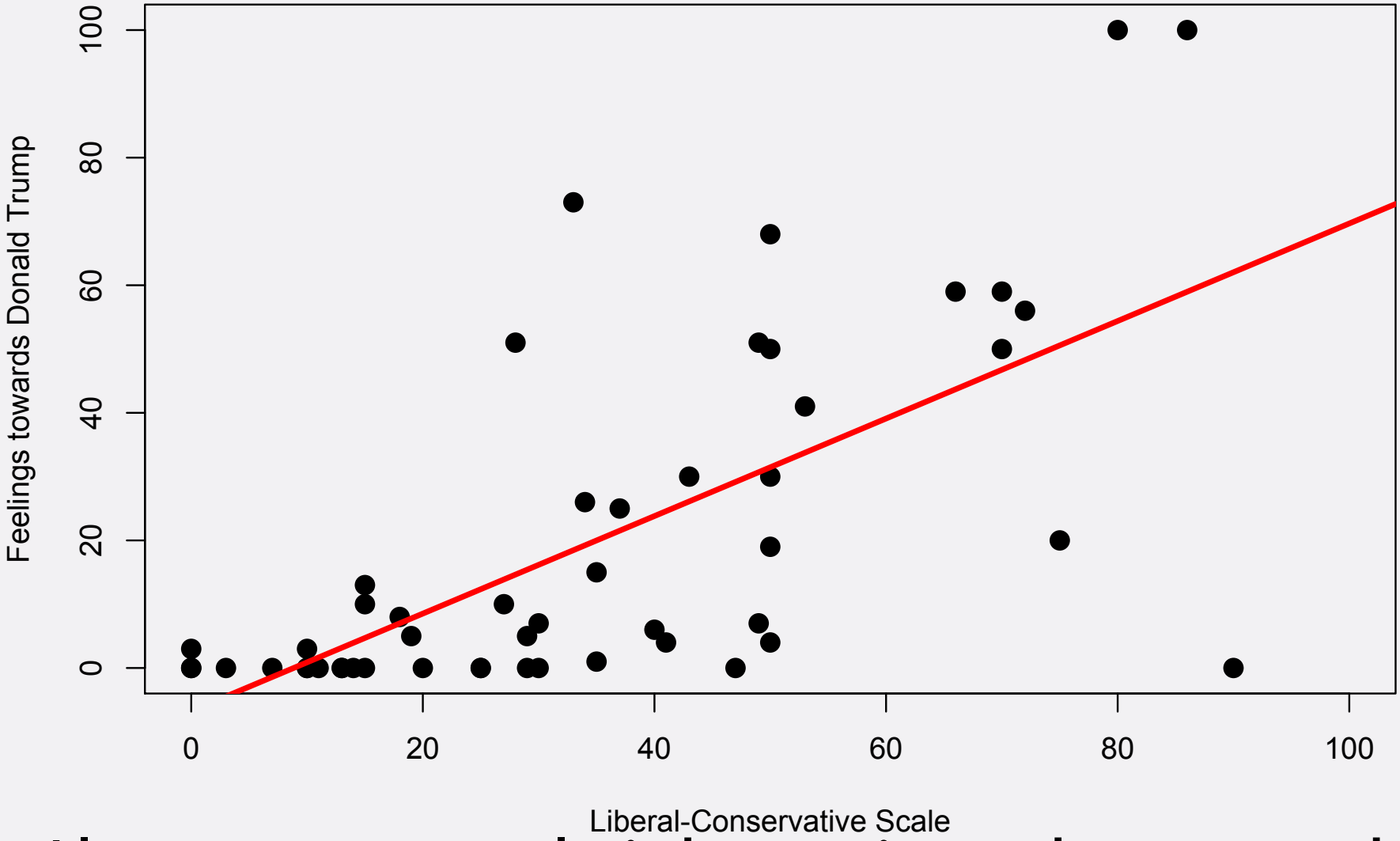
- Linear regression really means *linear*
- Often, effect of x on y is *not* linear

FROM OUR SURVEY



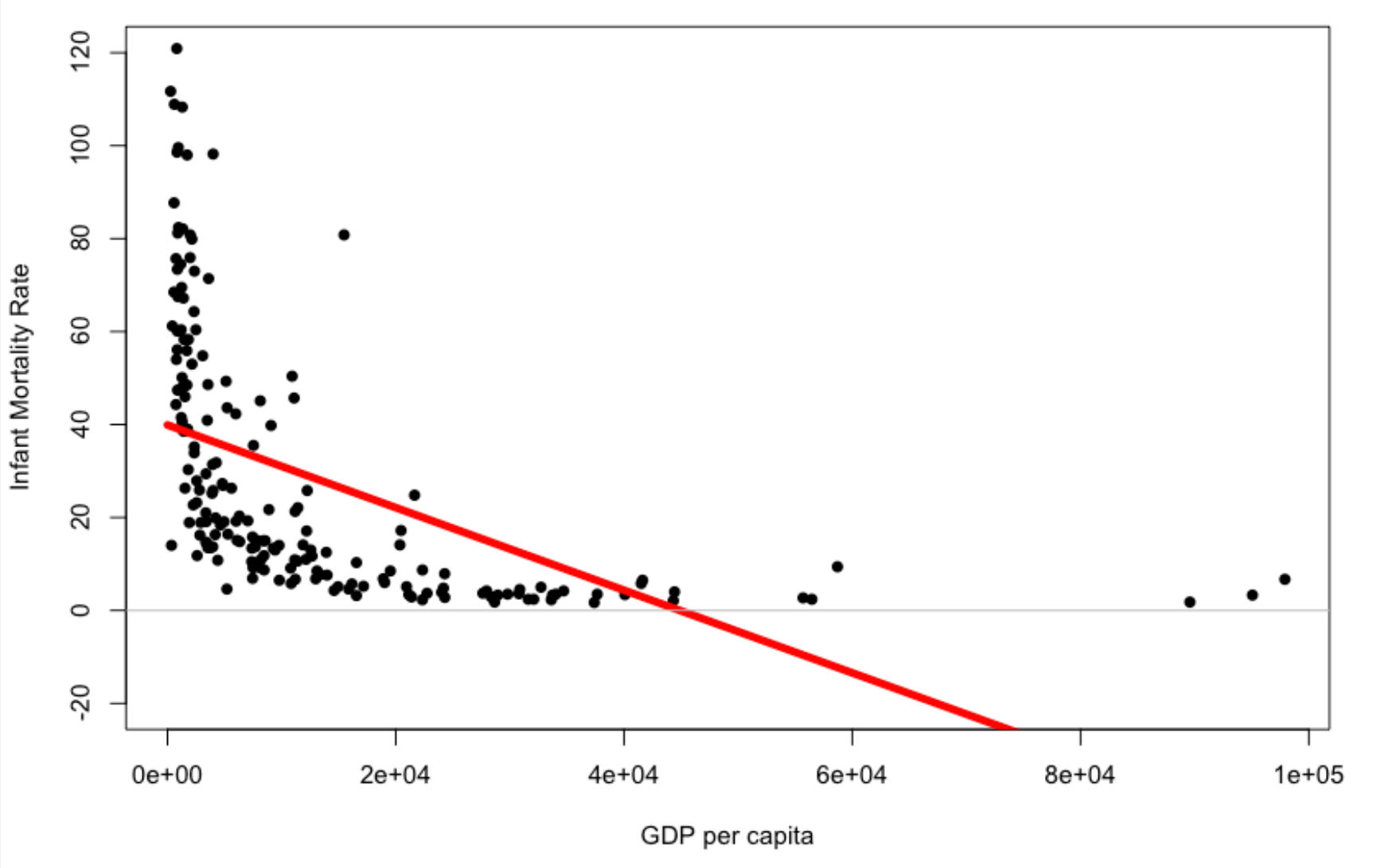
- **Score=-6.8 + 0.8 * Lib/Cons**
- **Intercept is negative!**

FROM OUR SURVEY



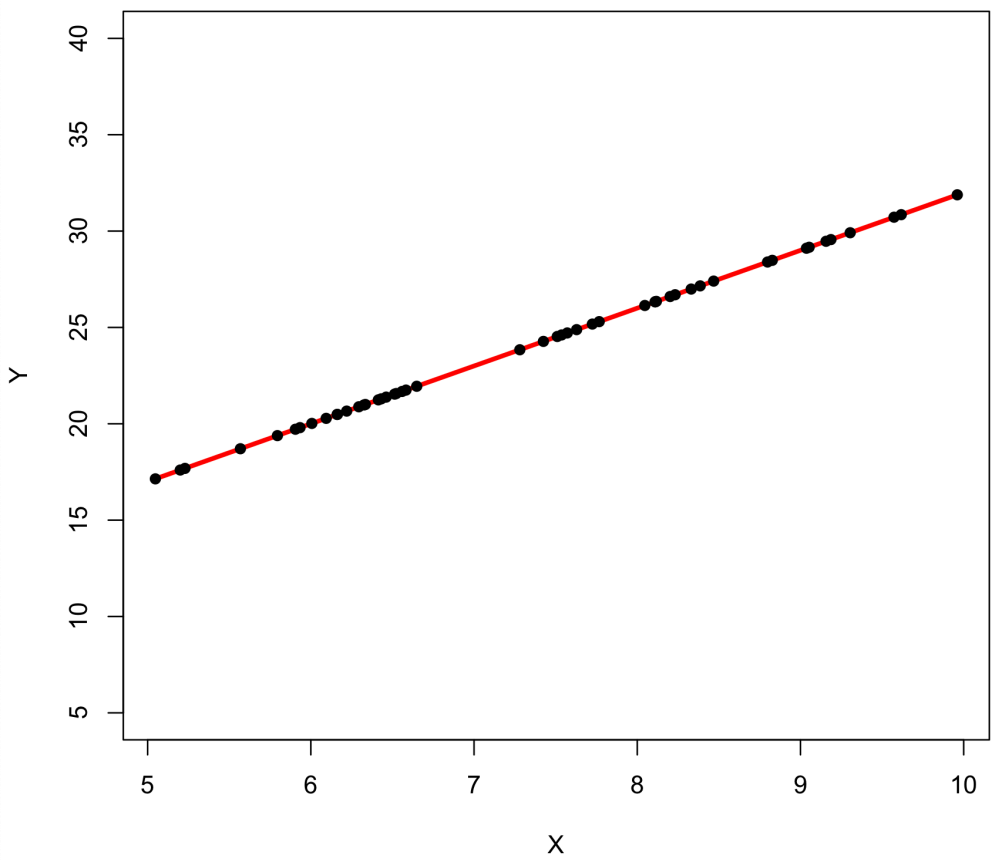
- **Always start an analysis by getting to know your data, make plots etc.**

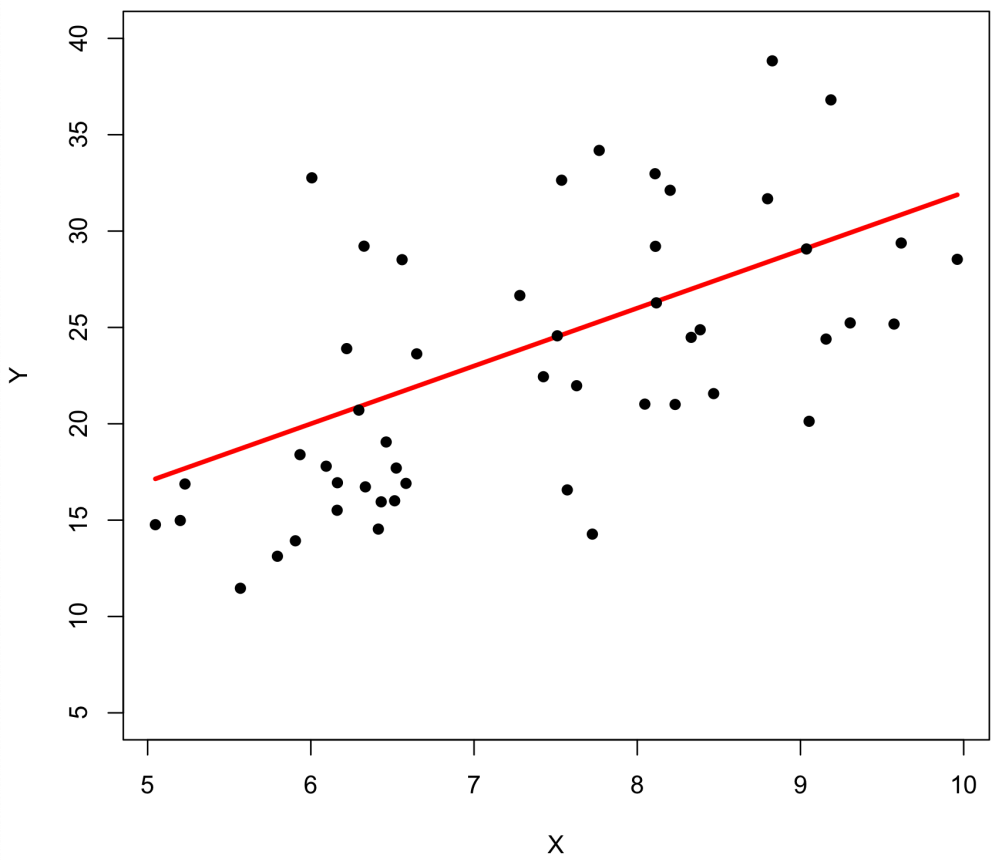
ANOTHER THING



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

MORE GENERALLY



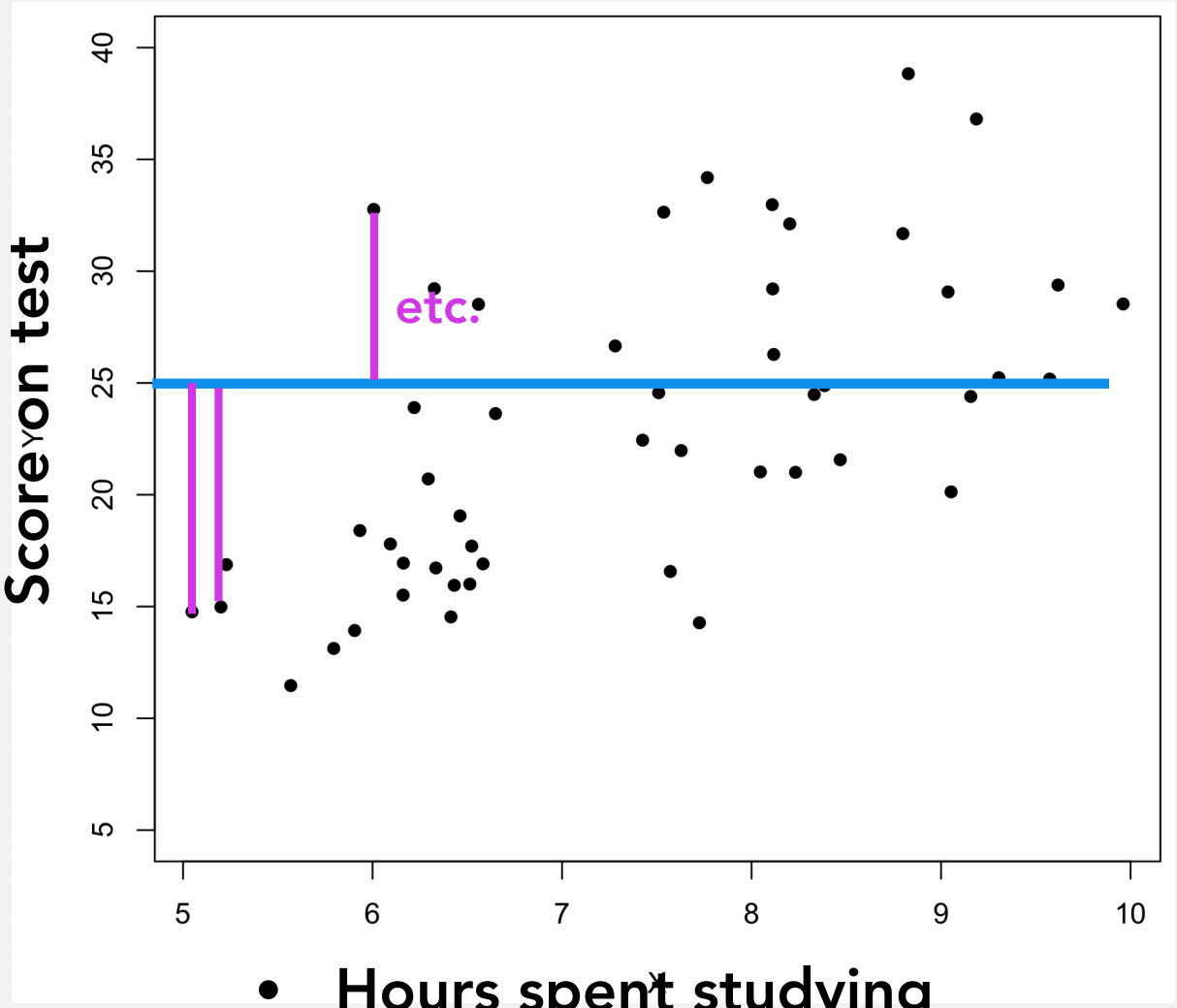


- Same regression equation in both situations
  - Y = 2 + 3*X
- But: X explains Y much better in the first than in the second
- Regression equation does not tell us *how much* it explains

EXPLANATORY POWER MEASURE

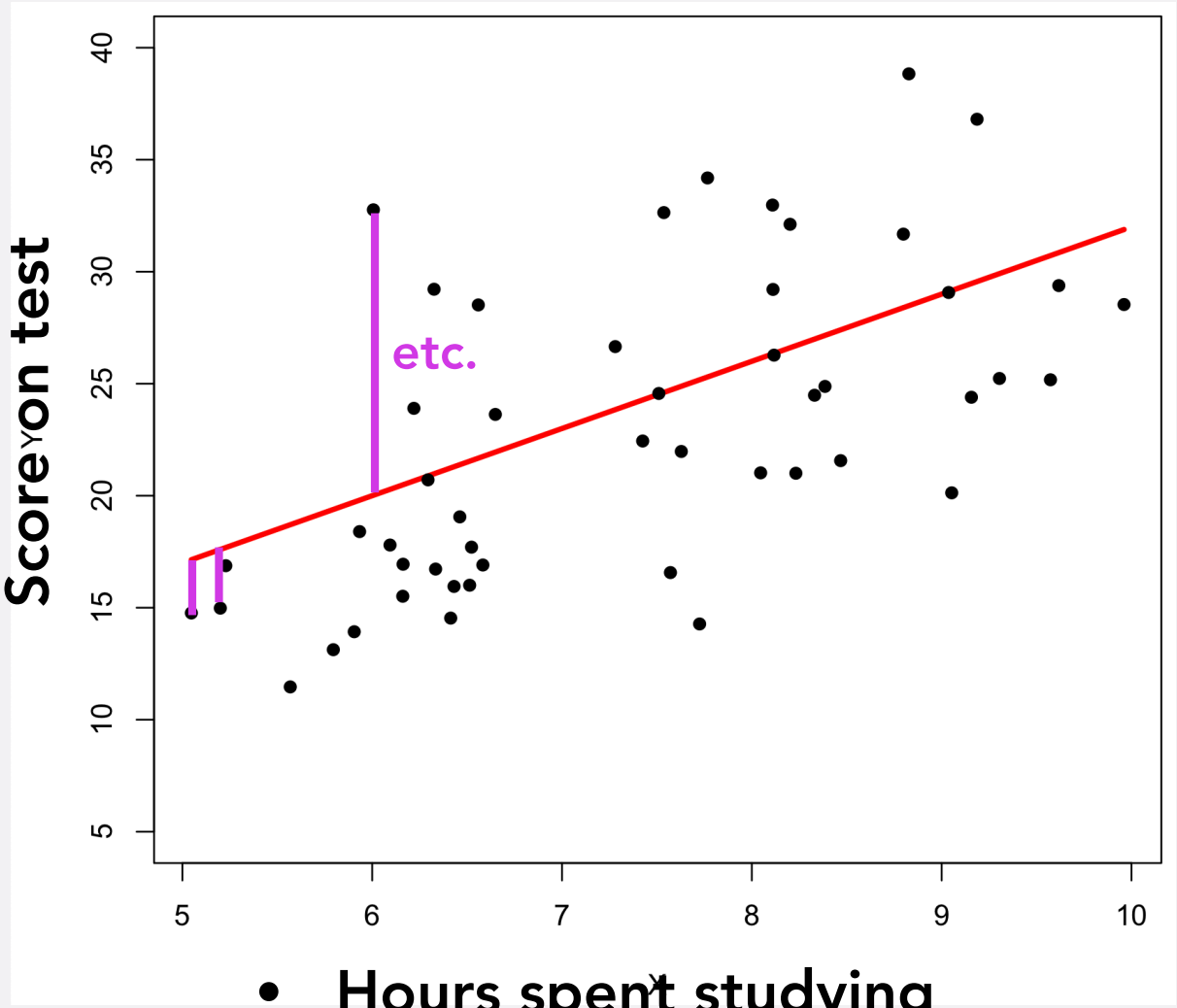
- Need: measure of how well independent variable explains dependent variable in a linear regression
- Idea: How much of the variation in Y can we predict using X?

EXPLANATORY POWER MEASURE



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

EXPLANATORY POWER MEASURE



- 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

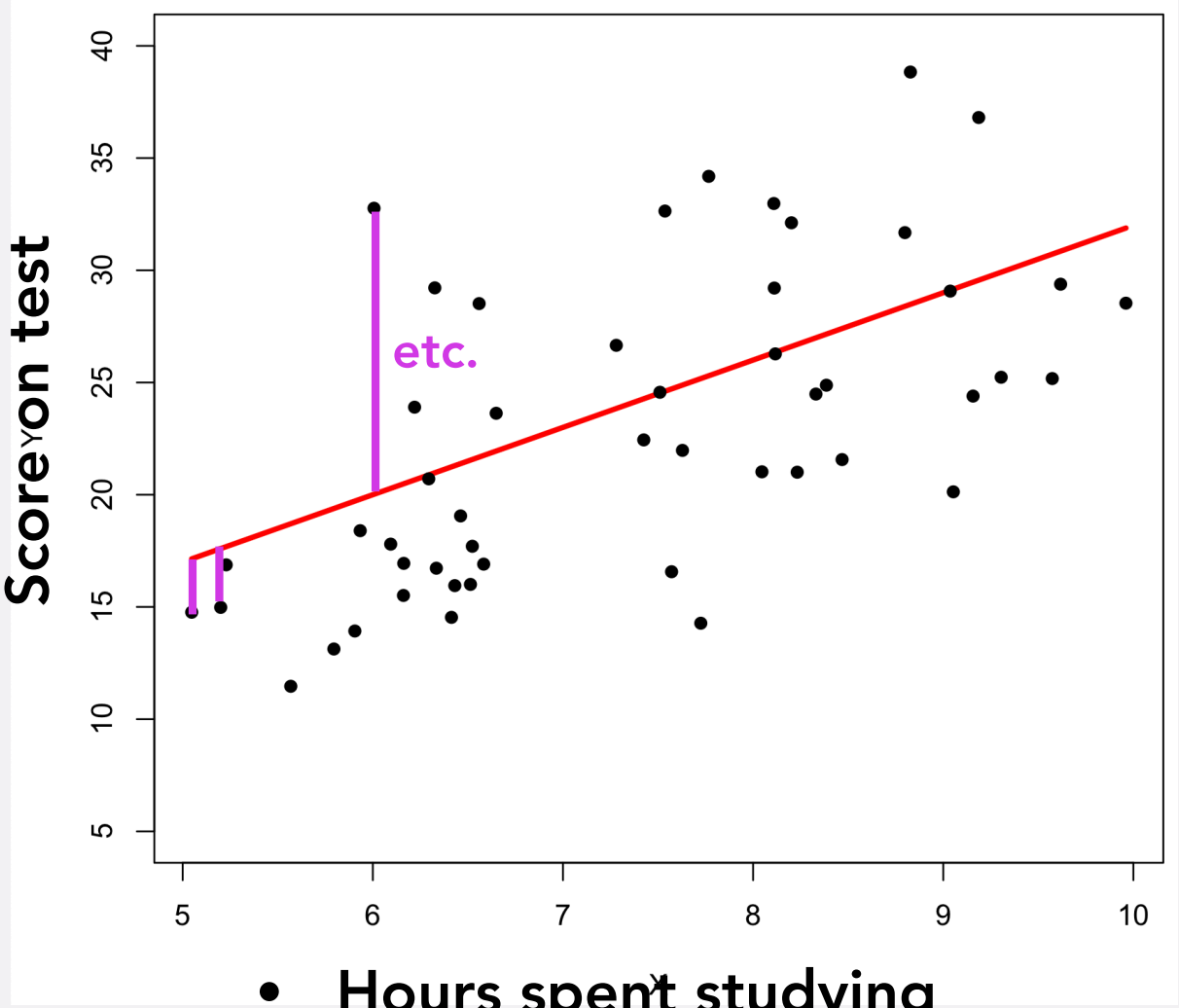
R-SQUARE

- 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

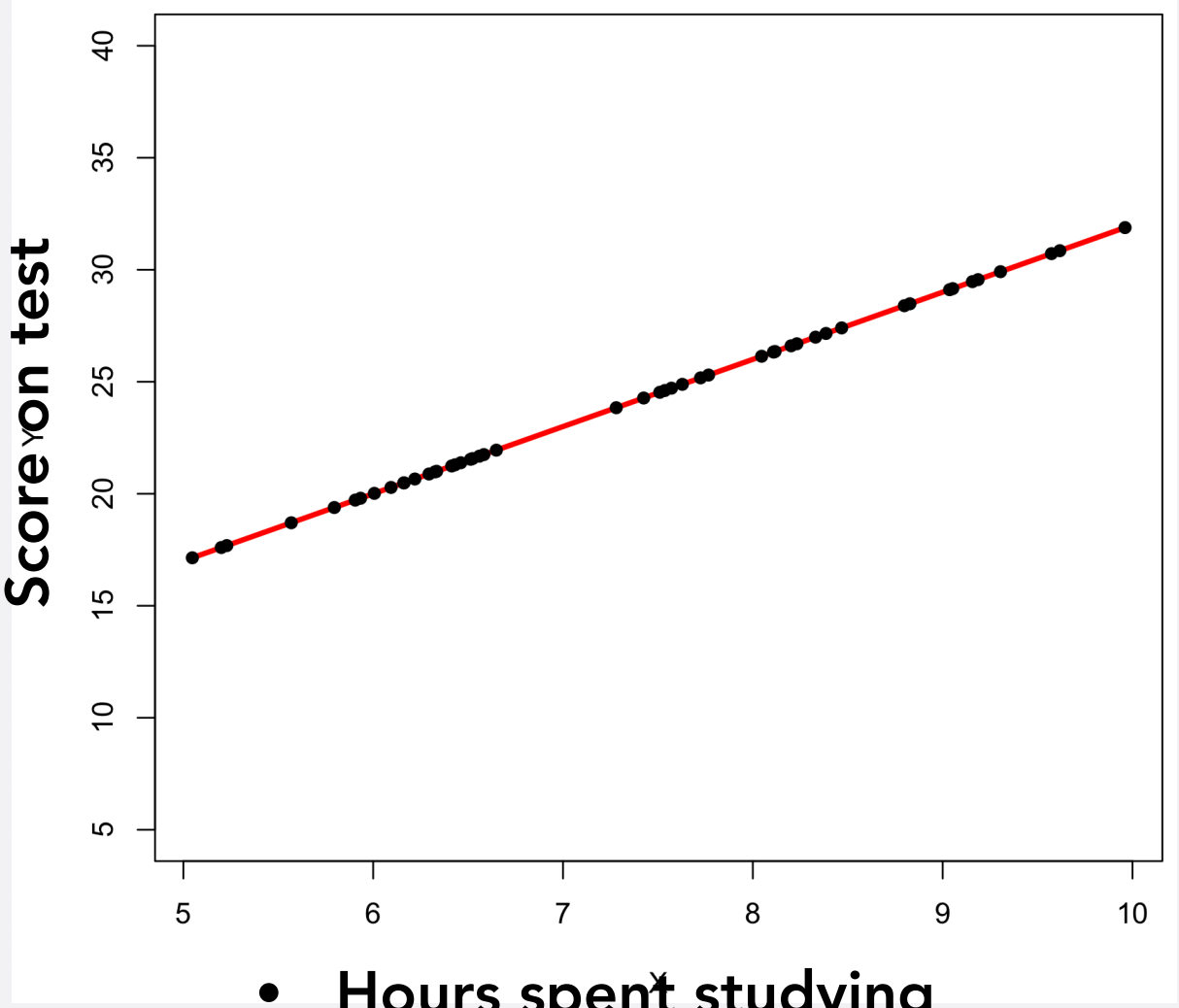
- Typically, not expressed as percentage (between 0 and 100), but as fraction (between 0 and 1)
  - 0: The independent variable explains *none* of the variation in the dependent variable
  - 1: The independent variable explains *all* of the variation in the dependent variable

EXPLANATORY POWER MEASURE



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

EXPLANATORY POWER MEASURE



- Hours spent studying explains 100% of variance in test scores
- So: $R^2 = 1$

BACK TO OUR EXAMPLE

```
. reg therm_2 libcons_1
```

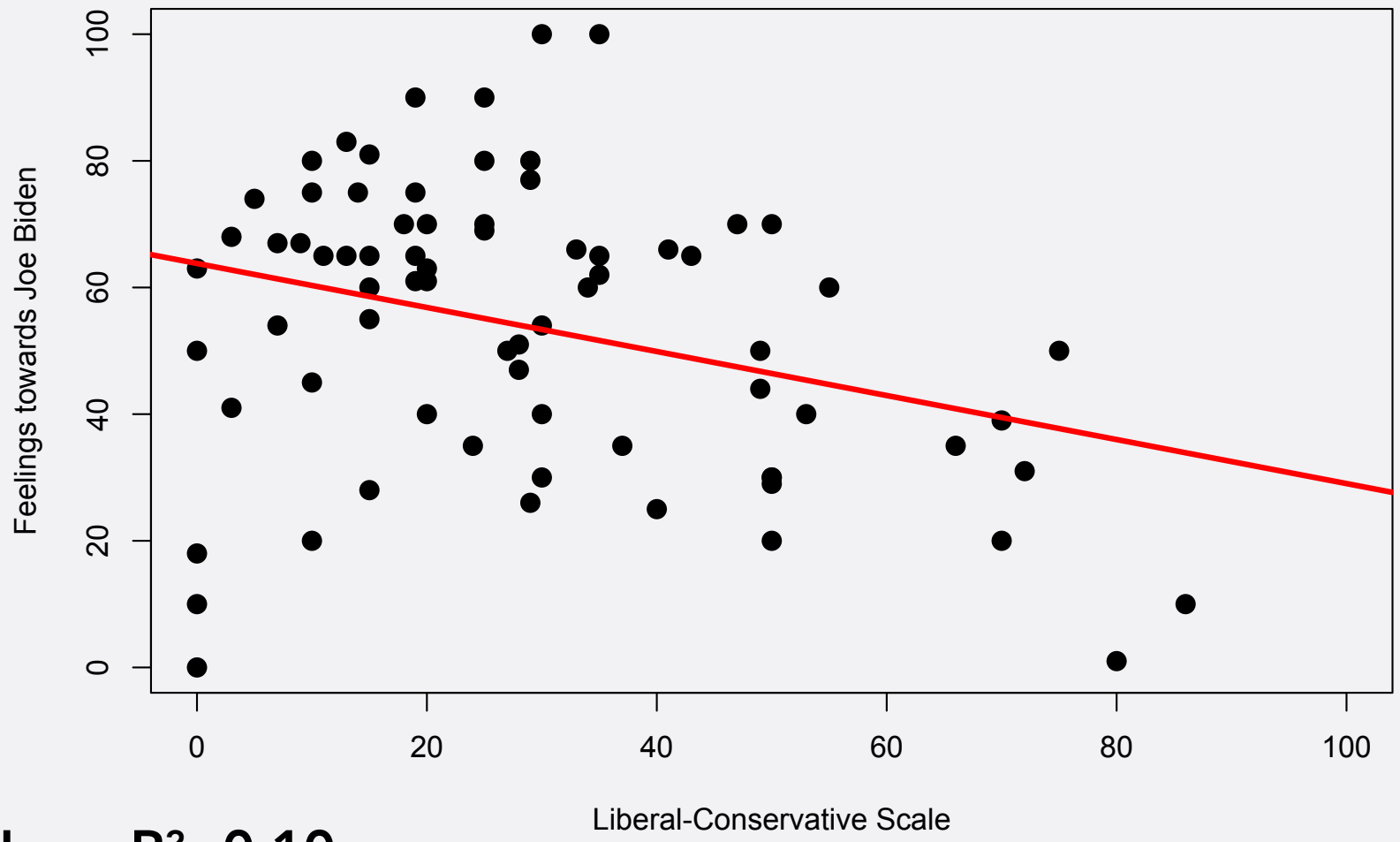
Source	SS	df	MS		
Model	3834.01698	1	3834.01698	Prob > F	0.0059
Residual	34232.5776	72	475.452467	R-squared	0.1007
Total	38066.5946	73	521.4602	Root MSE	21.805

	Number of obs =	74
	F(1, 72) =	8.06
	Adj R-squared =	0.0882

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

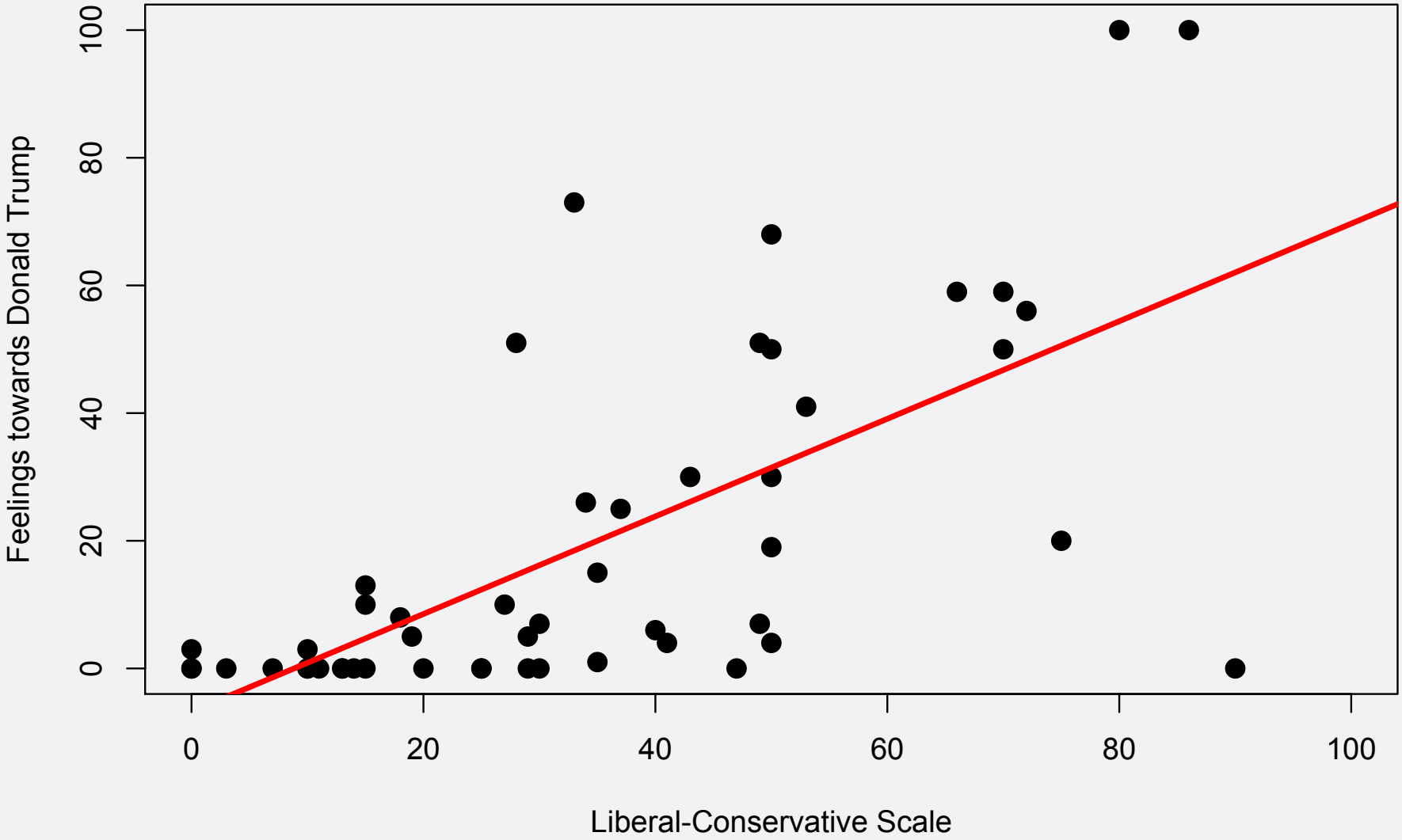
- DV: Rating of J. Biden (therm_2)
- IV: Liberal-conservative scale (libcons_1)

JOE BIDEN



- Here: R^2=0.10
- Lib/cons rating only explains 10% of variance in ratings of J. Biden
- So 90% remain unexplained…

DONALD TRUMP



DONALD TRUMP

```
reg therm_1 libcons_1
```

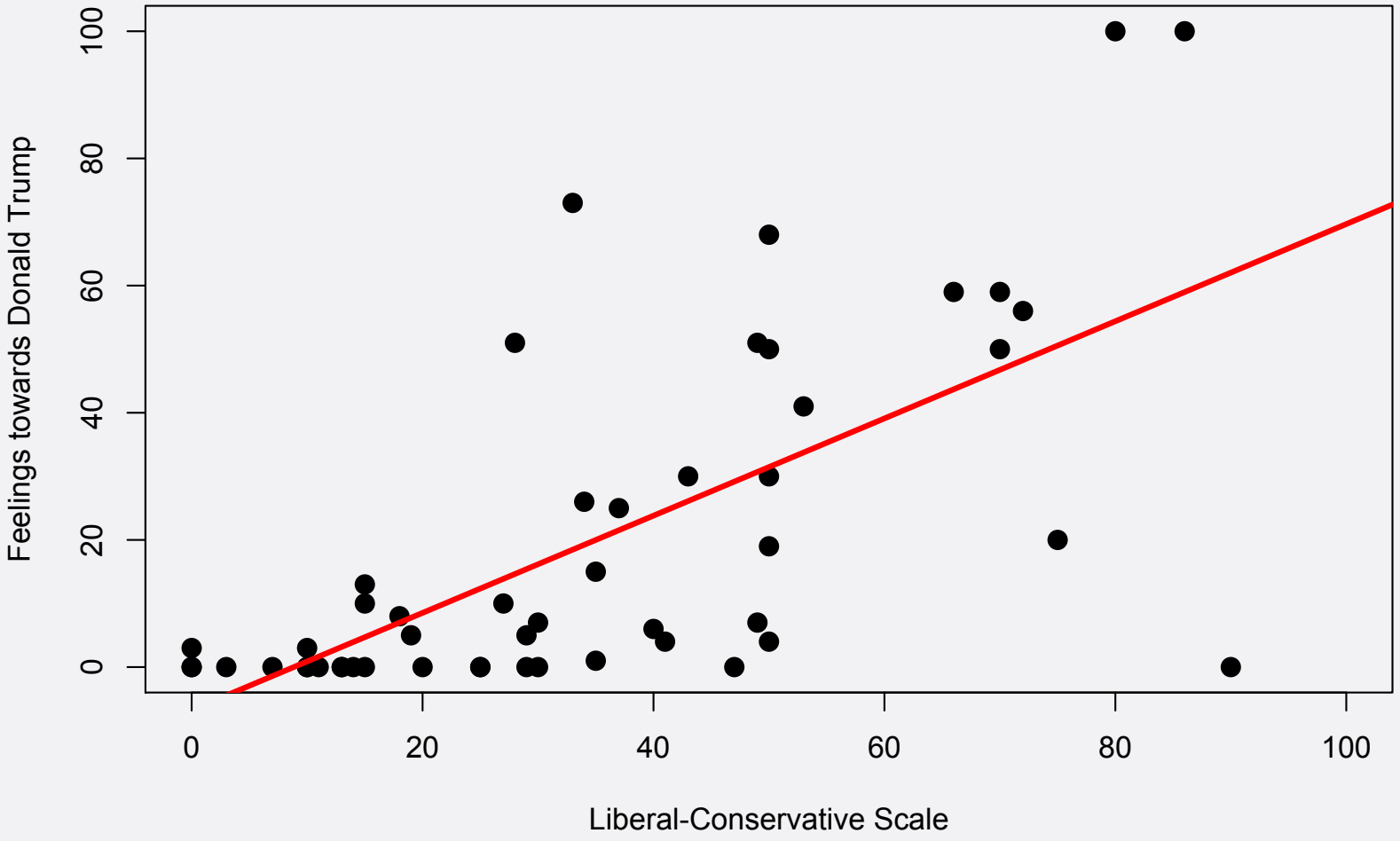
Source	SS	df	MS		
Model	16298.2825	1	16298.2825	Prob > F	[illegible]
Residual	20428.345	49	416.905	R-squared	0.4438
Total	36726.6275	50	734.532549	Root MSE	20.418

Number of obs	=	51				
F(1, 49)	=	39.09				
Prob > F	=	[illegible]				
R-squared	=	0.4438				
Adj R-squared	=	0.4324				
Root MSE	=	20.418				

therm_1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
libcons_1	.7643888	.1222537	6.25	0.000	.5187108	1.010067
_cons	-6.759461	5.118334	-1.32	0.193	-17.04514	3.526216

- DV: Rating of D. Trump (therm_1)
- IV: Liberal-conservative scale (libcons_1)

DONALD TRUMP

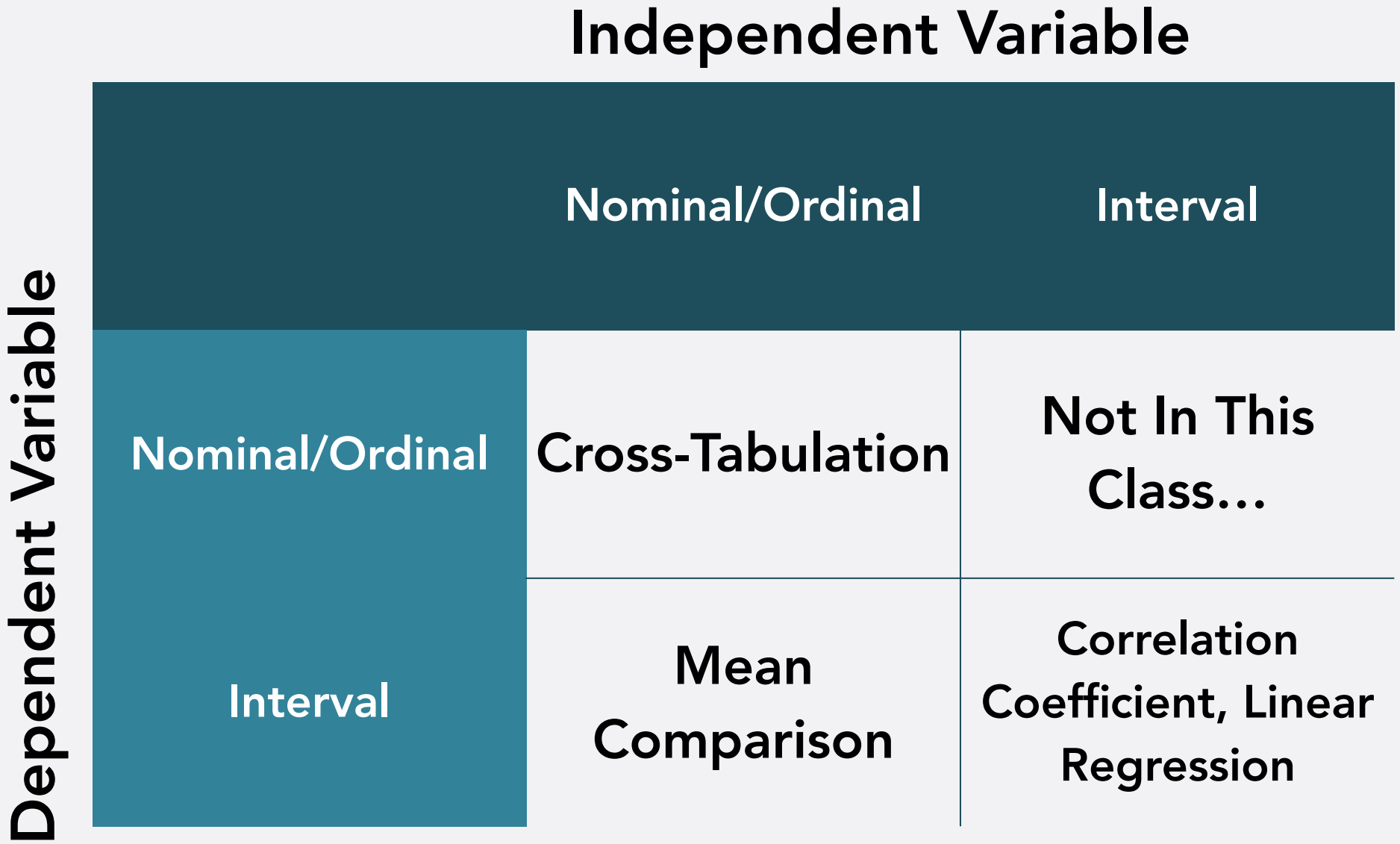


- Here: R^2=0.44
- Lib/cons rating explains 44% of variance in ratings of D. Trump
- So only 56% remain unexplained…

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



NOW

- 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?

REMEMBER

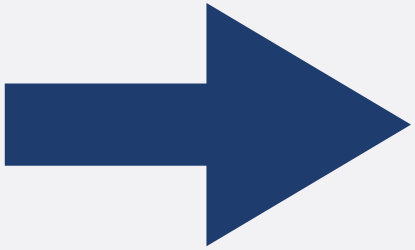
POLITICS SEPTEMBER 22, 2021

Biden's Approval Rating Hits New Low of 43%; Harris' Is 49%

Results for this Gallup poll are based on telephone interviews conducted Sept. 1-17, 2021, with a random sample of 1,005 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±4 percentage points at the 95% confidence level. All reported margins of sampling error include computed design effects for weighting.

BIVARIATE RELATIONSHIP

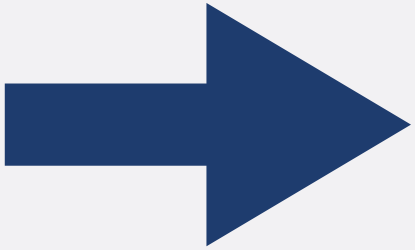
?



Approval for J. Biden

- What explains why some people approve of J. Biden while others do not?

BIVARIATE RELATIONSHIP

Gender  Approval for J. Biden

- Hypothesis: In a comparison of individuals, women are more likely to approve of J. Biden than men
  - "gender gap"

BIVARIATE RELATIONSHIP

Biden Approval Ratings Diverge by Gender, Education, Race

Job Approval Ratings of President Biden, by Subgroup

	Approve	Disapprove	N
	%	%	
All U.S. adults	56	39	2,937
Gender			
Men	49	45	1,643
Women	62	34	1,294

PROBLEM

- Is the effect of gender on approval real?
  - Does it exist in the population?
- Or is it only something that we found in our sample, but gender actually has no effect in the population?

PROBLEM

- We have a *random sample*
  - Men: 49% approval
  - Women: 62% approval
- Want to know: is mean approval rating of men and women in the *population* the same or not?

SOLUTION

- Idea: Use relation between two variables in *sample* to make inference about relation between two variables in *population*
  - Of course, means we can make mistakes

ALTERNATIVE HYPOTHESIS

- There *is* a relationship between the independent and dependent variable in the population

- H_A or H_1

NULL HYPOTHESIS

- In the population, there is *no relationship* between dependent and independent variable
  - If there is a difference in the sample, it is due to random sampling error

- H_0

IN OUR CASE

- H_0: In a comparison of individuals, there is *no difference* between men and women in approval of Biden
- H_A: In a comparison of individuals, women are more likely to approve of Biden than men

BACK TO MISTAKES

- Idea: Use relation between two variables in *sample* to make inference about relation between two variables in *population*
  - Of course, means we can make mistakes

ERRORS

	There Is A Relation In The Population	There Is No Relation In The Population
We Conclude There Is A Relation	✔	✘
We Conclude There Is No Relation	✘	✔

ERRORS

	There Is A Relation In The Population	There Is No Relation In The Population
We Conclude There Is A Relation	✔	✘
We Conclude There Is No Relation	✘	✔

TYPE I ERROR

- We conclude there is a relationship between X and Y when in reality there is not
  - "Type I error"
  - We falsely reject H_0

TYPE I ERROR

- We conclude there is a relationship between X and Y when in reality there is not
  - Example: There is no difference between men and women in approval rating in the population, but we conclude there is

ERRORS

	There Is A Relation In The Population	There Is No Relation In The Population
We Conclude There Is A Relation	✔	✘
We Conclude There Is No Relation	✘	✔

TYPE II ERROR

- We conclude there is no relationship between X and Y when in reality there is
  - "Type II error"
  - We falsely do not reject H_0

TYPE II ERROR

- We conclude there is no relationship between X and Y when in reality there is
  - Example: There is a difference between men and women in approval rating in the population, but we conclude there is none

ERRORS

	There Is A Relation In The Population	There Is No Relation In The Population
We Conclude There Is A Relation	✔	✘ Type I
We Conclude There Is No Relation	✘ Type II	✔

DECISION

- It's really bad if we conclude there is a relationship when in reality there is not (Type I error)
  - Type II error is also not great, but not as bad
- Thus: We privilege H_0

DECISION

- By default: We start out with assumption that there is no relationship in population (so H_0 is true)
  - No difference between men and women in Biden approval in population

DECISION

- Ask: Is there enough evidence in the *sample* to reject H_0?
  - Is the observed difference between mean and women in *sample* large enough to reject null hypothesis that no difference between them in population?