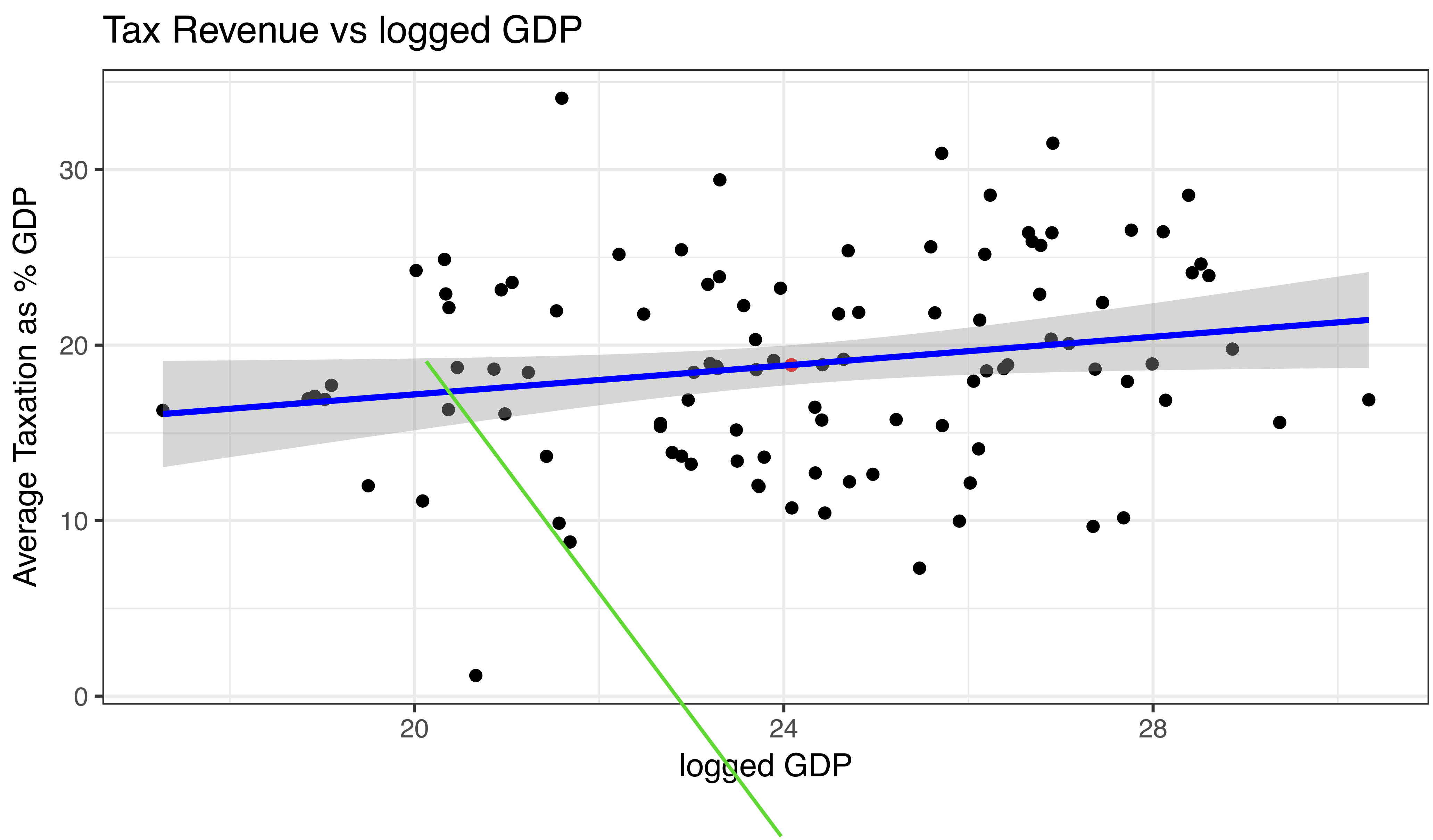
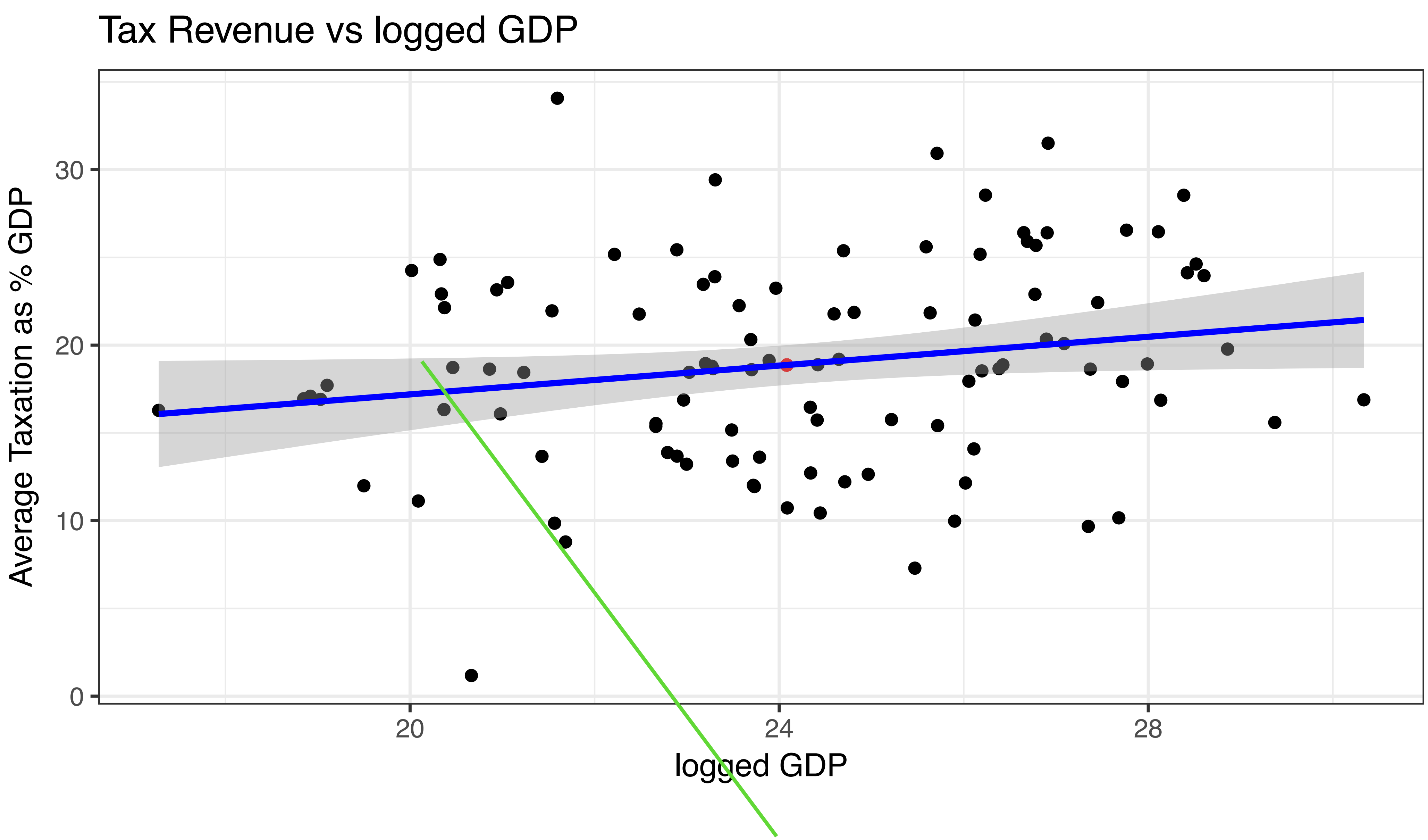
Regression lines minimize distance to all points

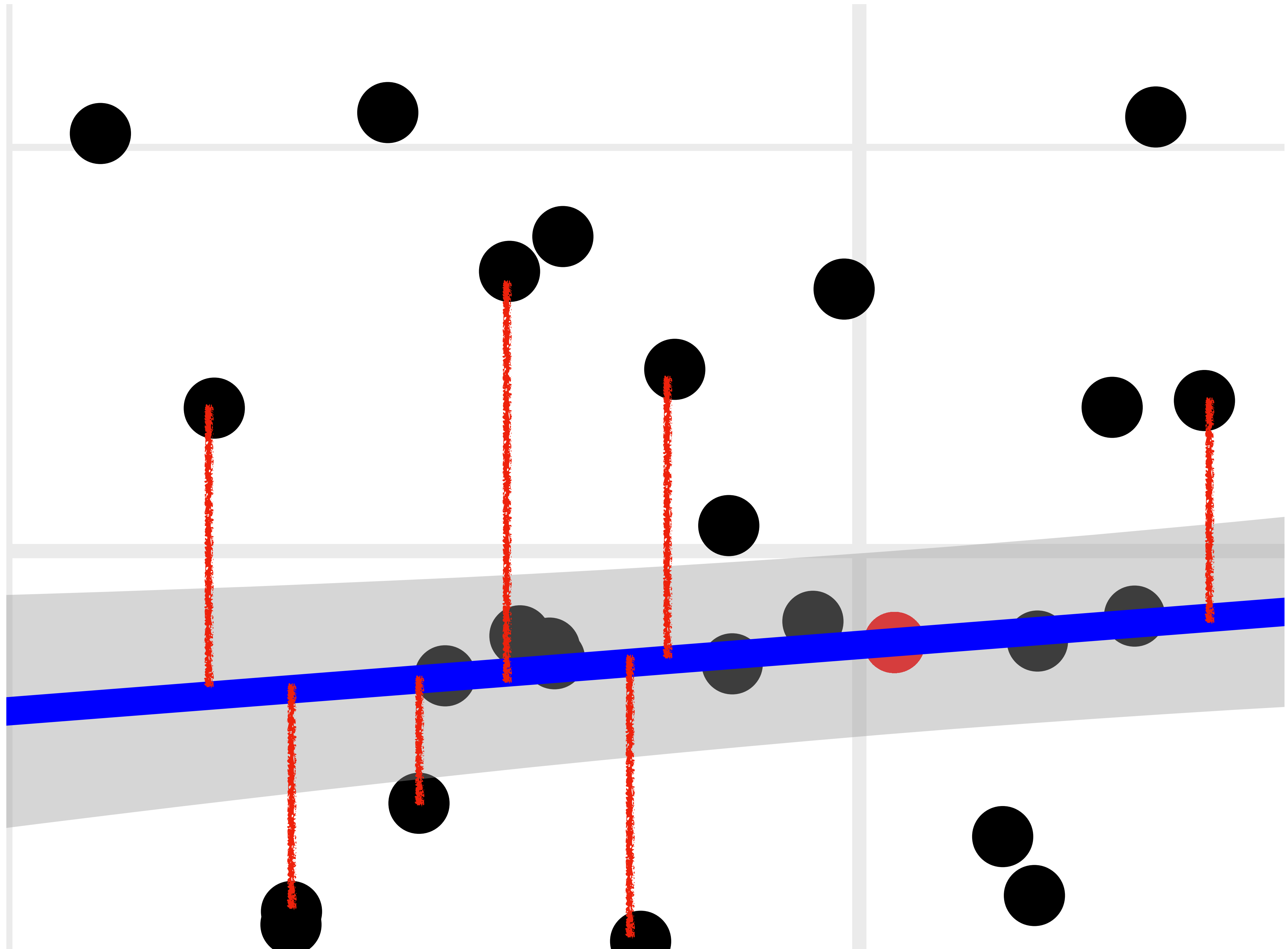


But the line does not go through all points

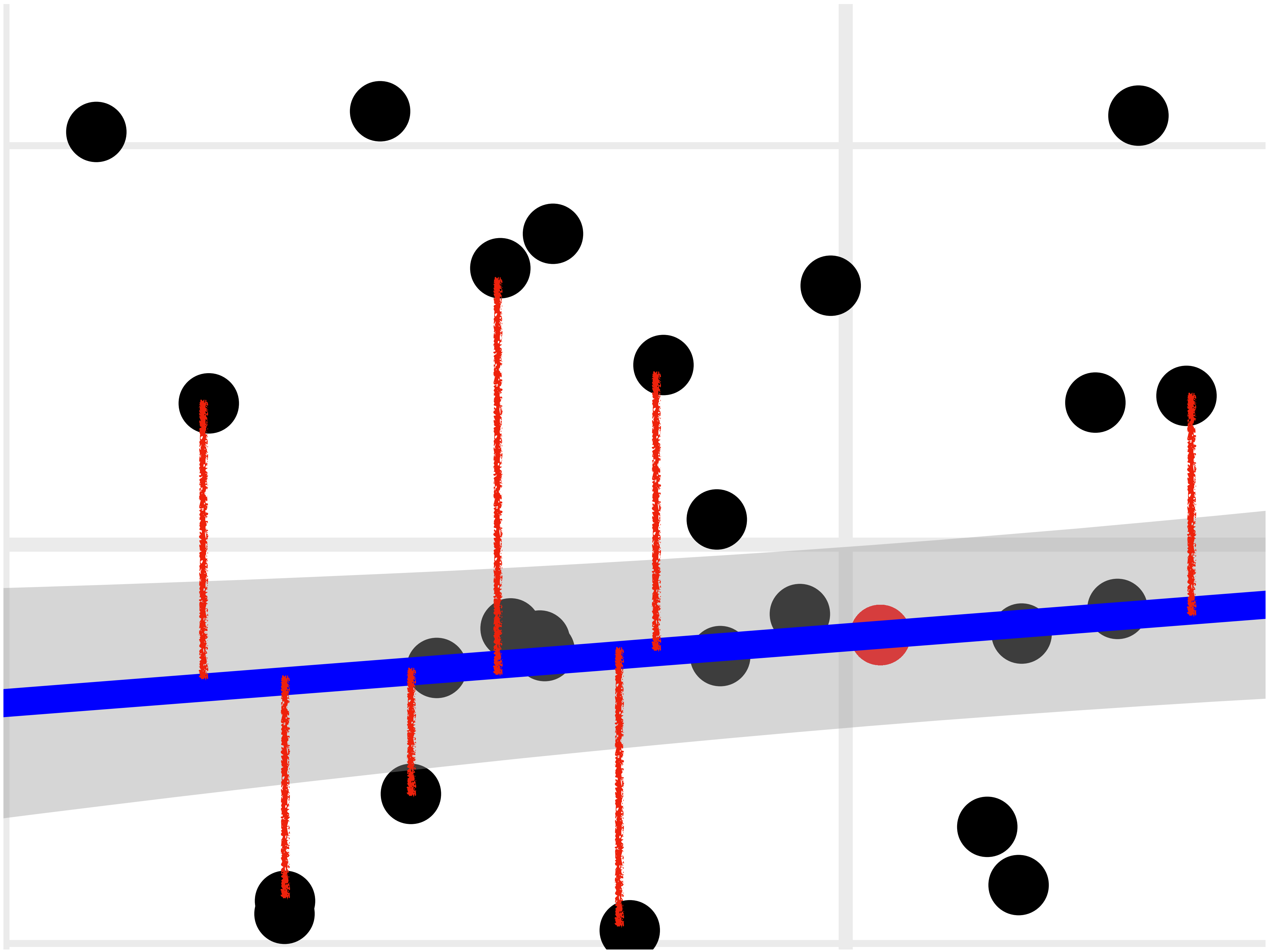


Each point is associated with an error:

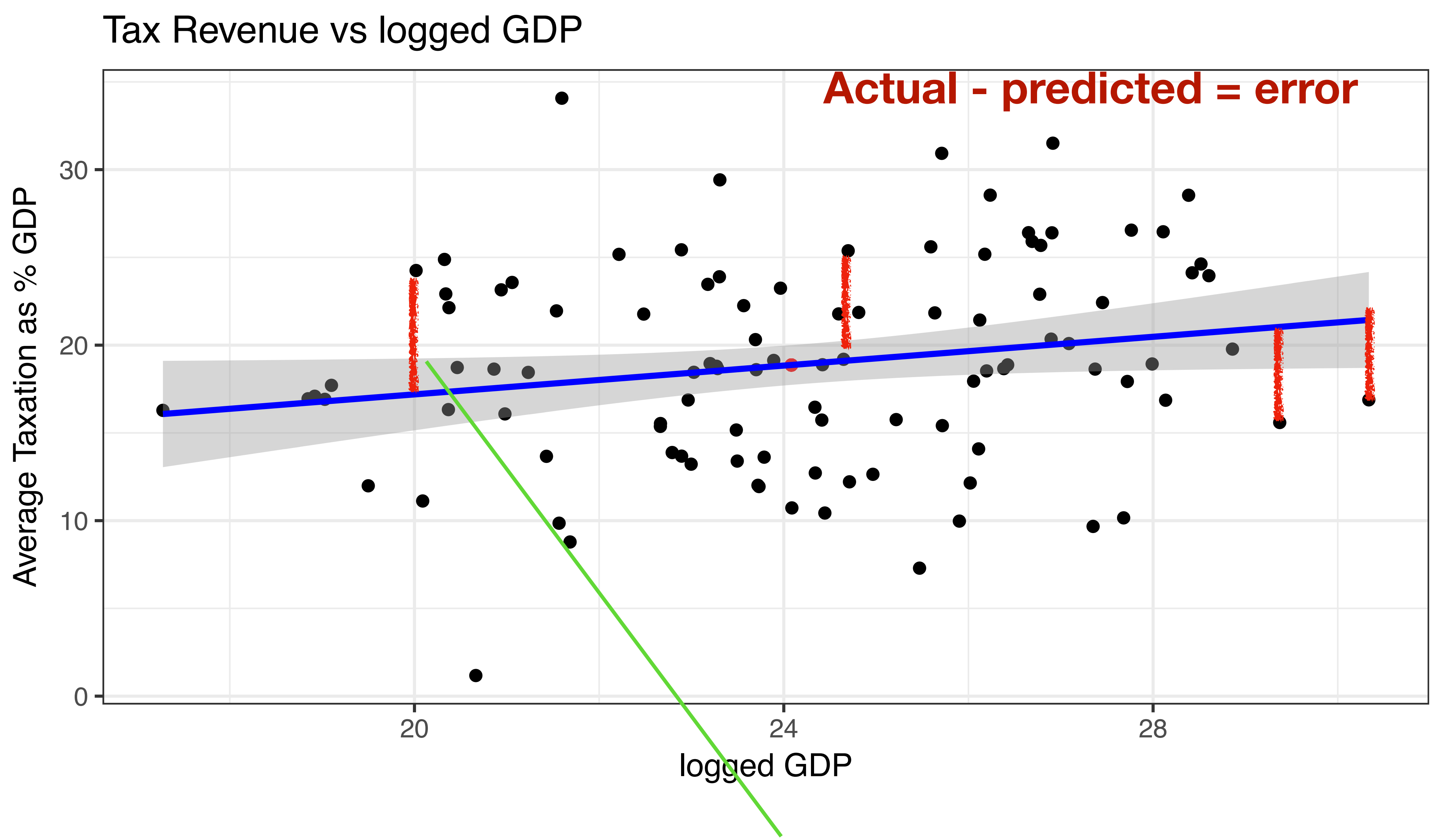
prediction at x - actual value of y at x



Error = prediction - actual y

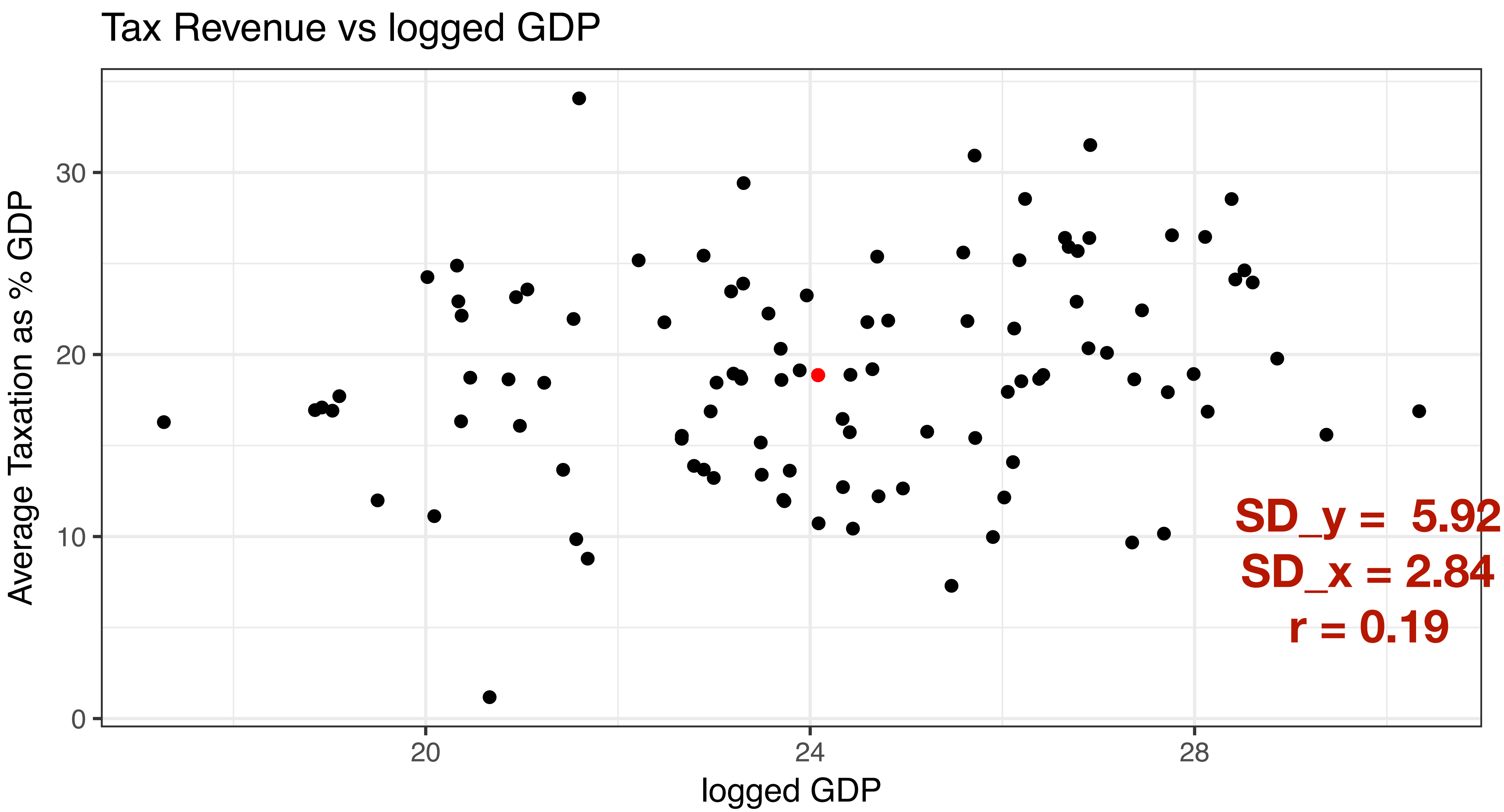


But regression lines are not perfect



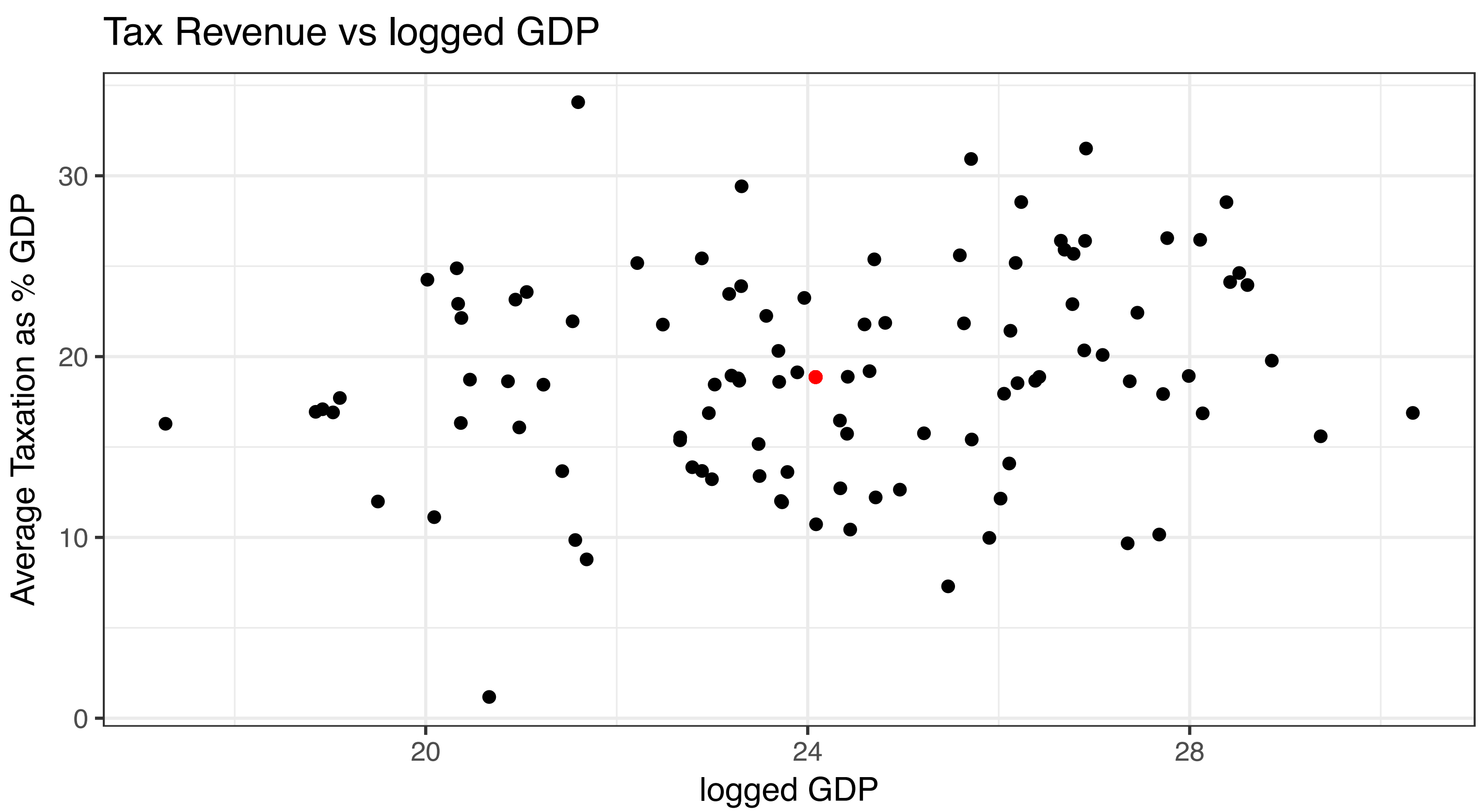
We always measure the error in terms of prediction error in y! Why?

Example of error calculation



Slope of regression line: 5.92*0.19/2.84 = 0.4

Example of error calculation

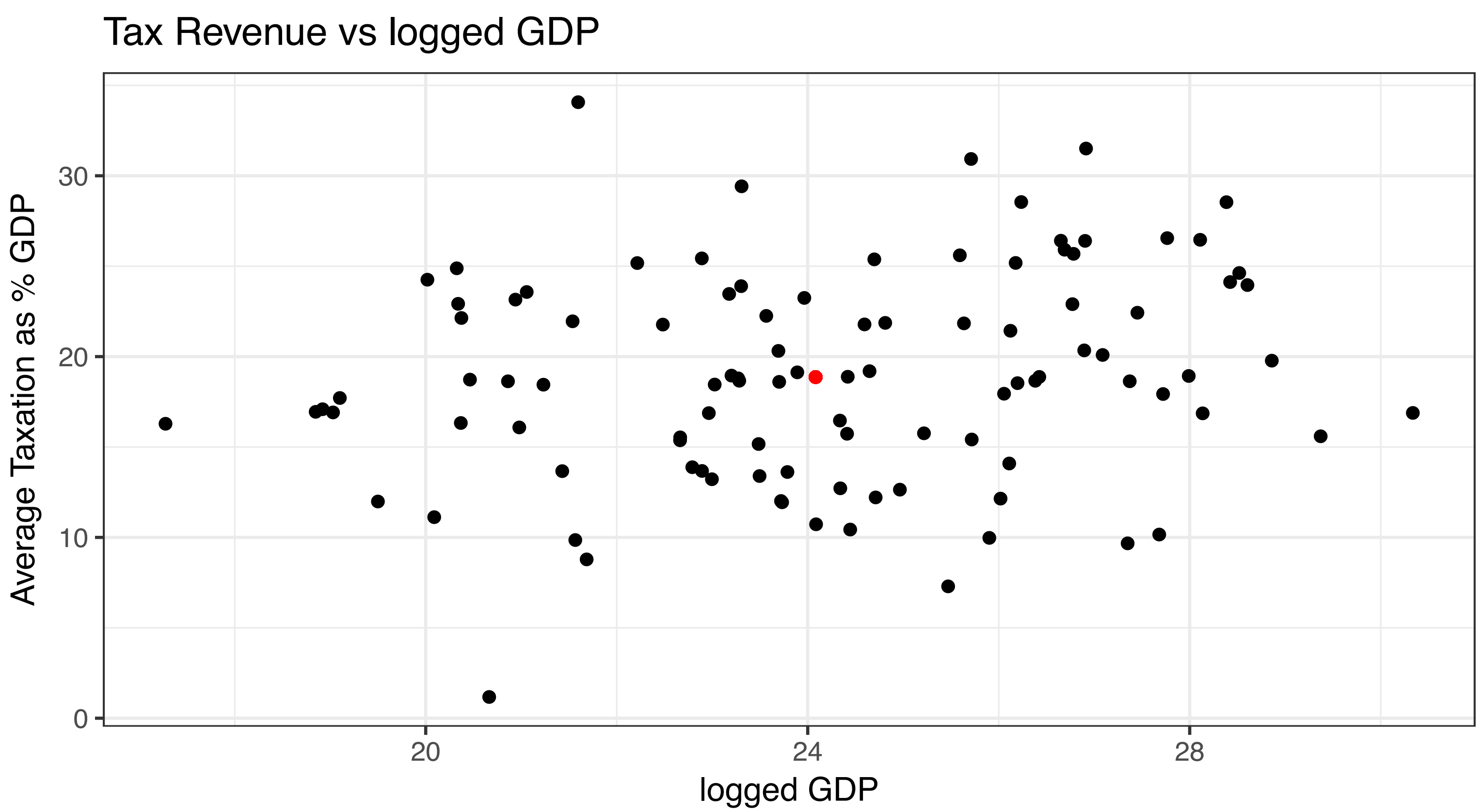


Mean_y = 18.87
Mean_x = 24.09
SD_y = 5.92
SD_x = 2.84
r = 0.19

Slope of regression line: 0.4

What is the prediction for x = 27.35????

Example of error calculation

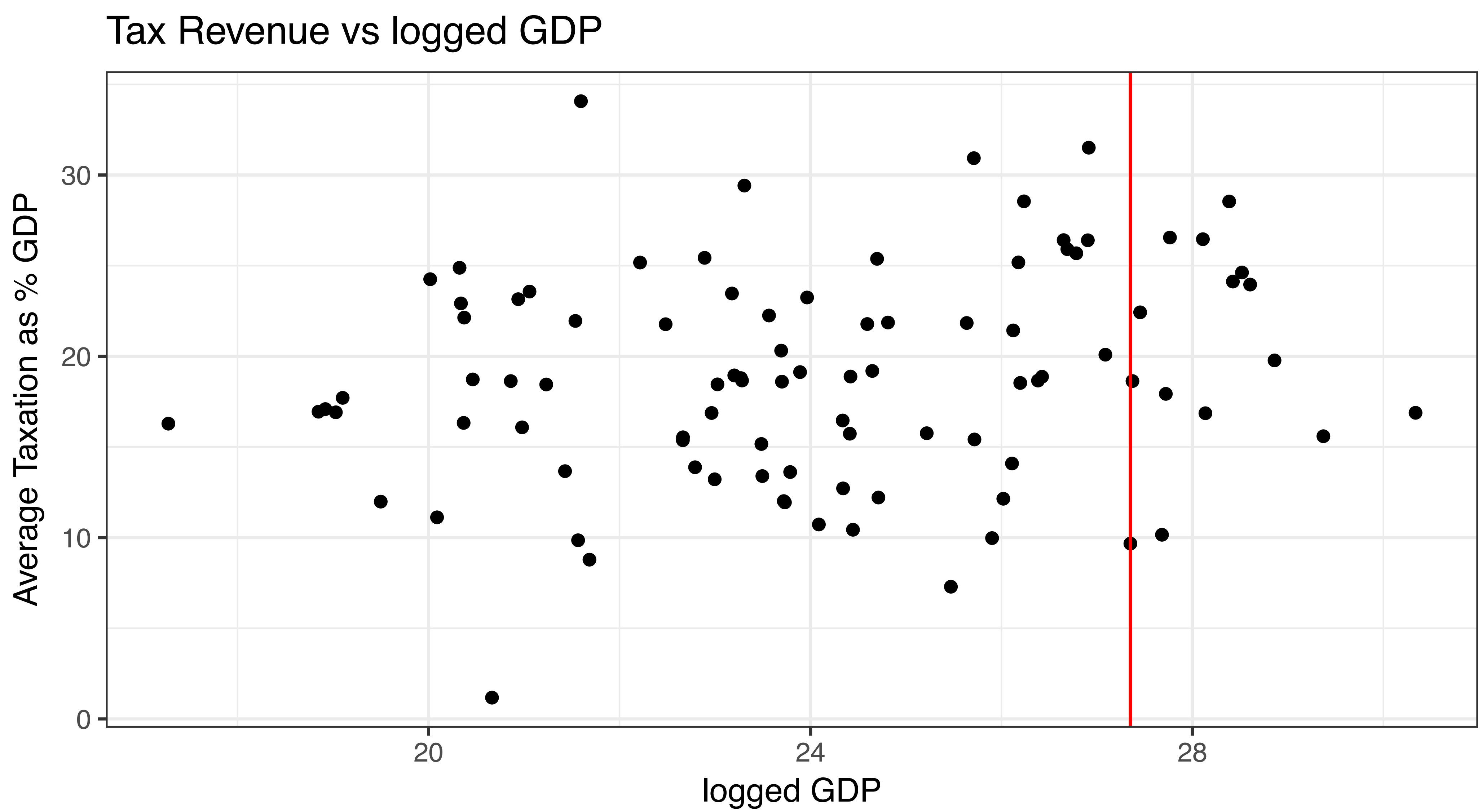


Mean_y = 18.87
Mean_x = 24.09
SD_y = 5.92
SD_x = 2.84
r = 0.19

Slope of regression line: 0.4

What is the prediction for x = 27.35????

Example of error calculation



What is the prediction for x = 27.35????
y_pred = 20.16
Actual Y: 9.67
Error = 9.67 - 20.16 = -10.49

Coach Sumlin asked for a prediction of the number of running plays that the Florida Gators will run on Saturday given that 2 inches of rain are expected. The correlation between rain in inches and number of running plays is 0.6. The average amount of rain in Gainesville is 0.5 inches with a standard deviation of 1. The Florida Gators run 35 running plays on average, with a standard deviation of 8.6. Based on 2 inches of rain, what is your prediction for the number of run plays executed by the Gators on Saturday?

But we had 0 inches of rain. What is the prediction?

32.42 predicted run plays

Actual number of run plays: 42

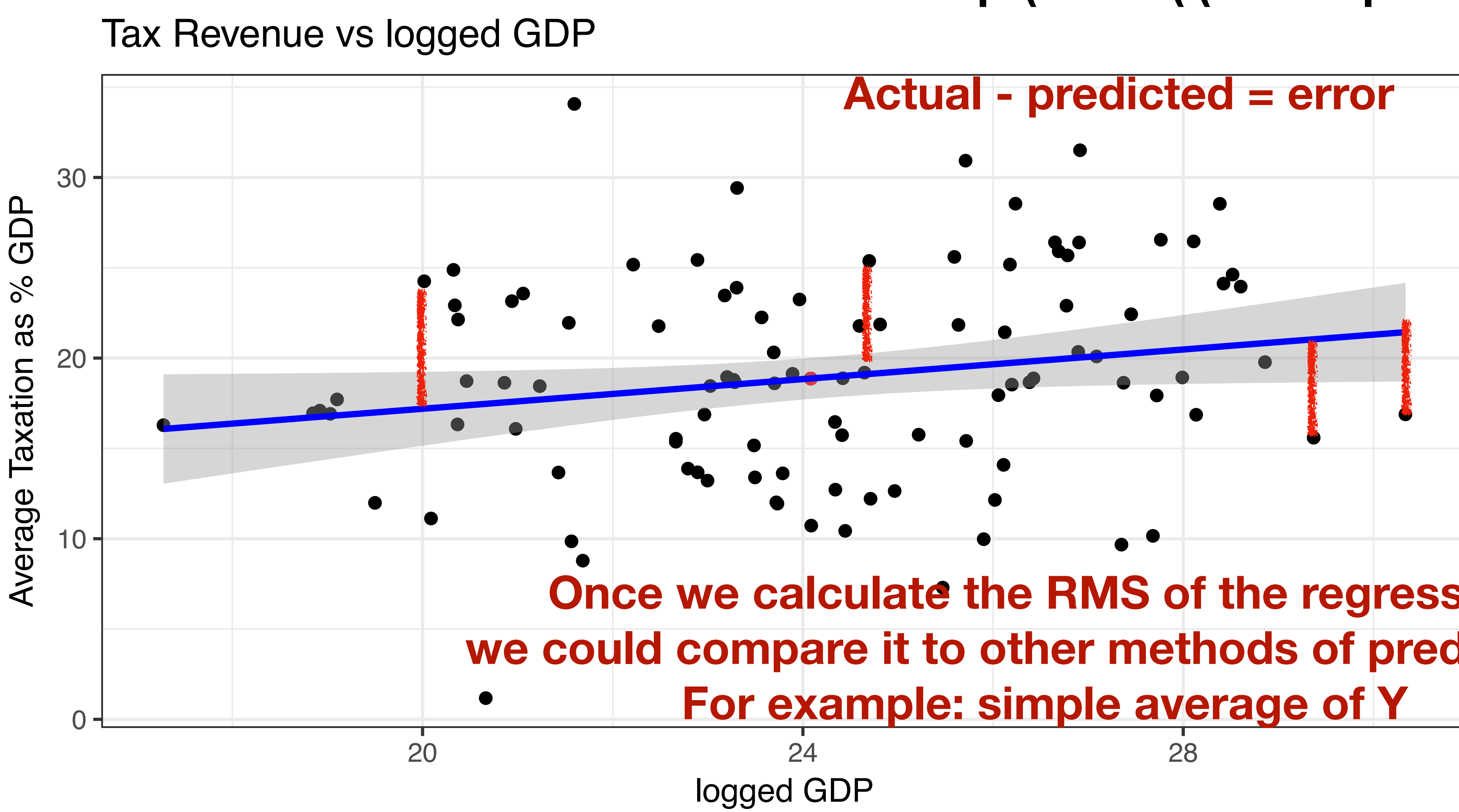
Error: 42 - 32.42 = 9.58

Recall the root mean squared error

- RMS = square root of the mean of the squared errors
- Approximately equal to the average of how far points are above and below the line
- RMS is always in the unit of the dependent variable (the variable to be predicted - y)
- Why can't we just take the average of the errors?

But regression lines are not perfect

RMS = sqrt(mean ((actual-predicted)^2))



Once we calculate the RMS of the regression,
we could compare it to other methods of predicting Y.
For example: simple average of Y

Recall the root mean squared error

- What is the root mean squared error of using the average of y to predict y?

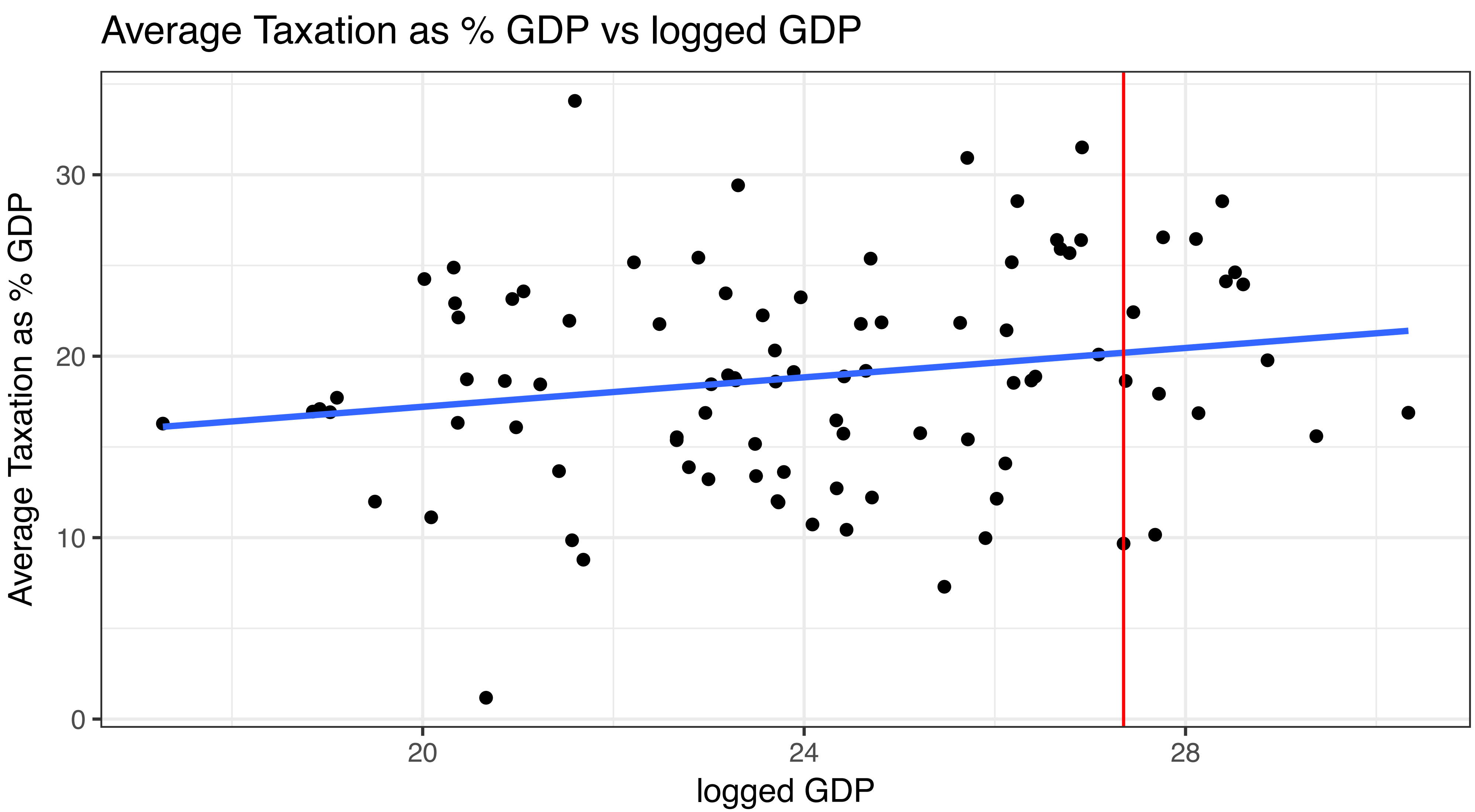
Recall the root mean squared error

- What is the root mean squared error of using the average of y to predict y?

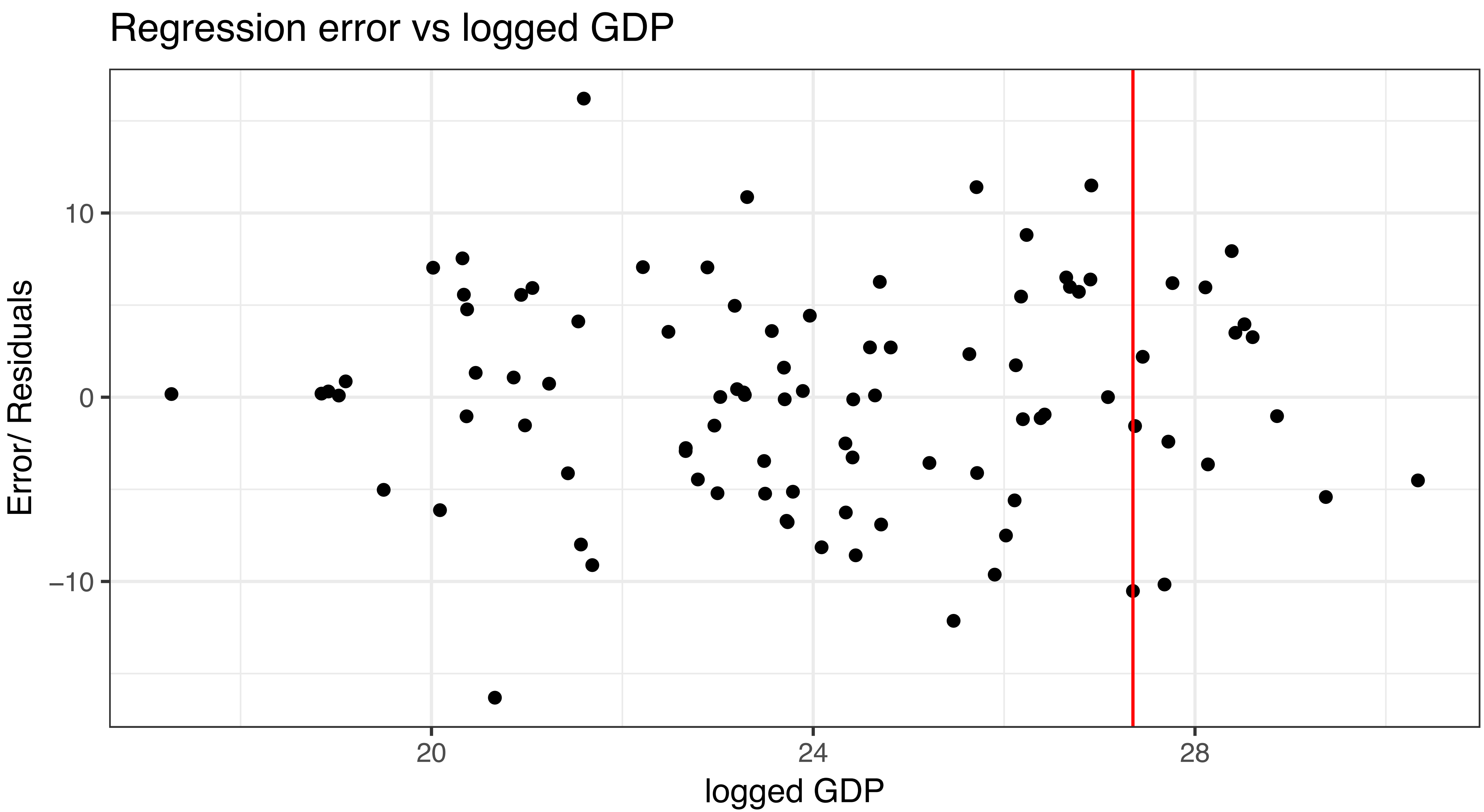
- The standard deviation!

Computing the rms for the regression

- In theory, we could calculate the rms by doing the calculation for every point in our data
- Luckily, we have a formula that makes calculation much simpler: rms_regression = SD_y * sqrt(1 - r^2)
- Again: rms is in the same units as the dependent variable
- In earlier example, rms would be in tax as % of GDP



Plotting Errors or Residuals



Plotting Errors or Residuals

Often the error is also called the residuals

- We can plot the error/residuals against the x-axis
- The residuals should average out to zero
- Regression line through residuals should be flat
- If residuals look funnel shaped, things are problematic

Homoscedasiticity

- Spread around the regression line is similar (the same) along the whole line
- The accuracy of predictions given the regression line should be the same along the whole line
- Football-shaped scatter plot
- If this condition is violated, we say the regression suffers from heteroscadasticity

Normal approximation in vertical strips

- What is the new average?
- What is the new SD?
- Everything else stays the same

Exercise

- Law school finds the following relationship btw. LSAT scores and first-year scores:
  - Average LSAT: 162, SD = 6
  - Average first-year score: 68, SD = 10,
  - R = 0.6

A. What is the percentage of students with first-year scores above 75?

B. Of students who scored 165 on LSAT, what percentage had first-year score greater than 75?

Exercise

- Law school finds the following relationship btw. LSAT scores and first-year scores:
  - Average LSAT: 162, SD = 6
  - Average first-year score: 68, SD = 10,
  - R = 0.6

A. What is the percentage of students with first-year scores above 75?

B. Of students who scored 165 on LSAT, what percentage had first-year score greater than 75?

Exercise

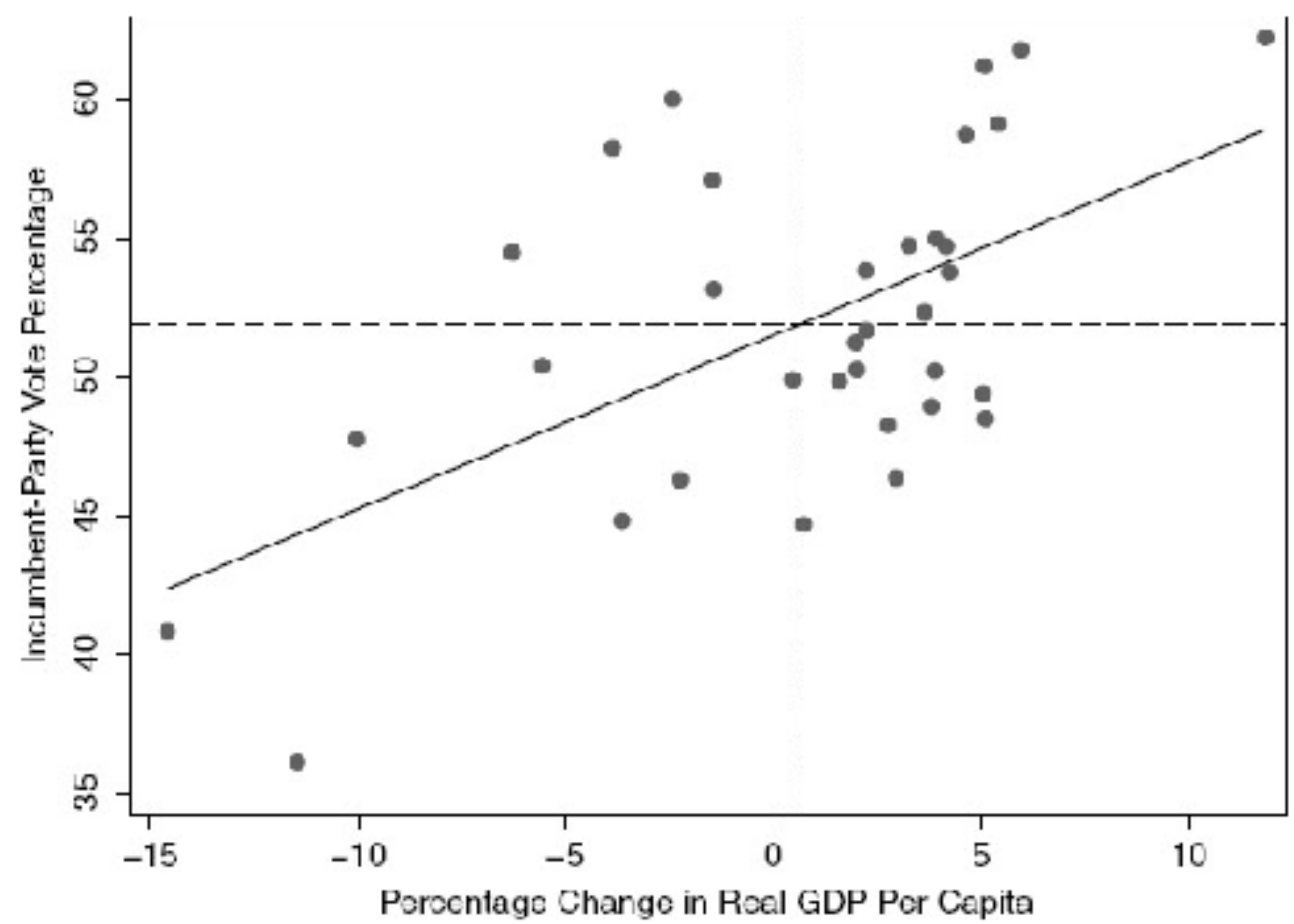
- Correlation in height for 66 boys:
  - Average height at 6, 3 feet and 10 inches, SD = 1.7 inches
  - Average height at 18, 5 feet and 10 inches, SD= 2.5
  - R = 0.8

A. RMS for regression predicting height at 18 from height at 6

B. RMS for regression predicting height at 6 from height at 18

The full regression line

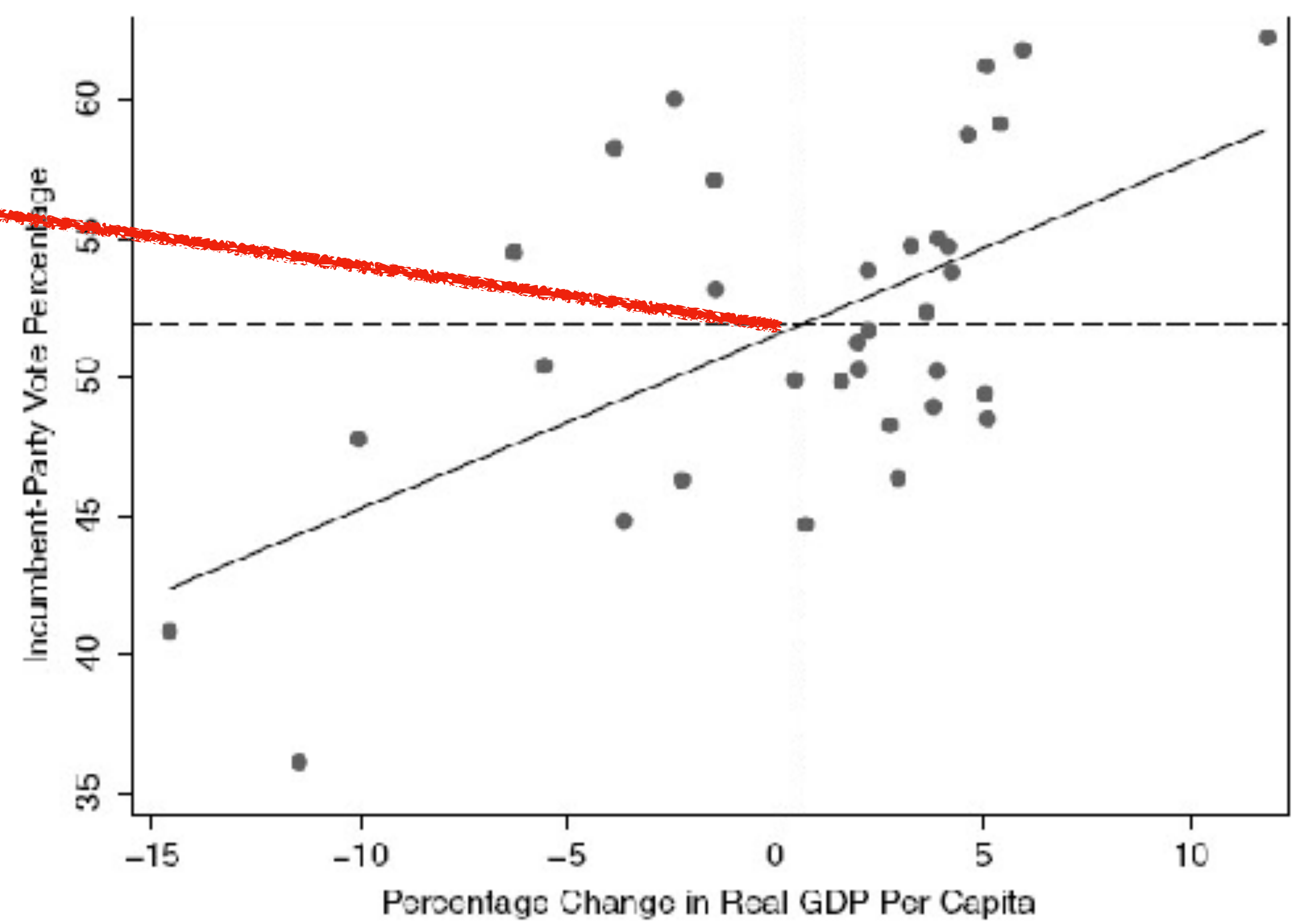
- Remember the formula of a line: $y = mx + b$
- So far we have only talked about m

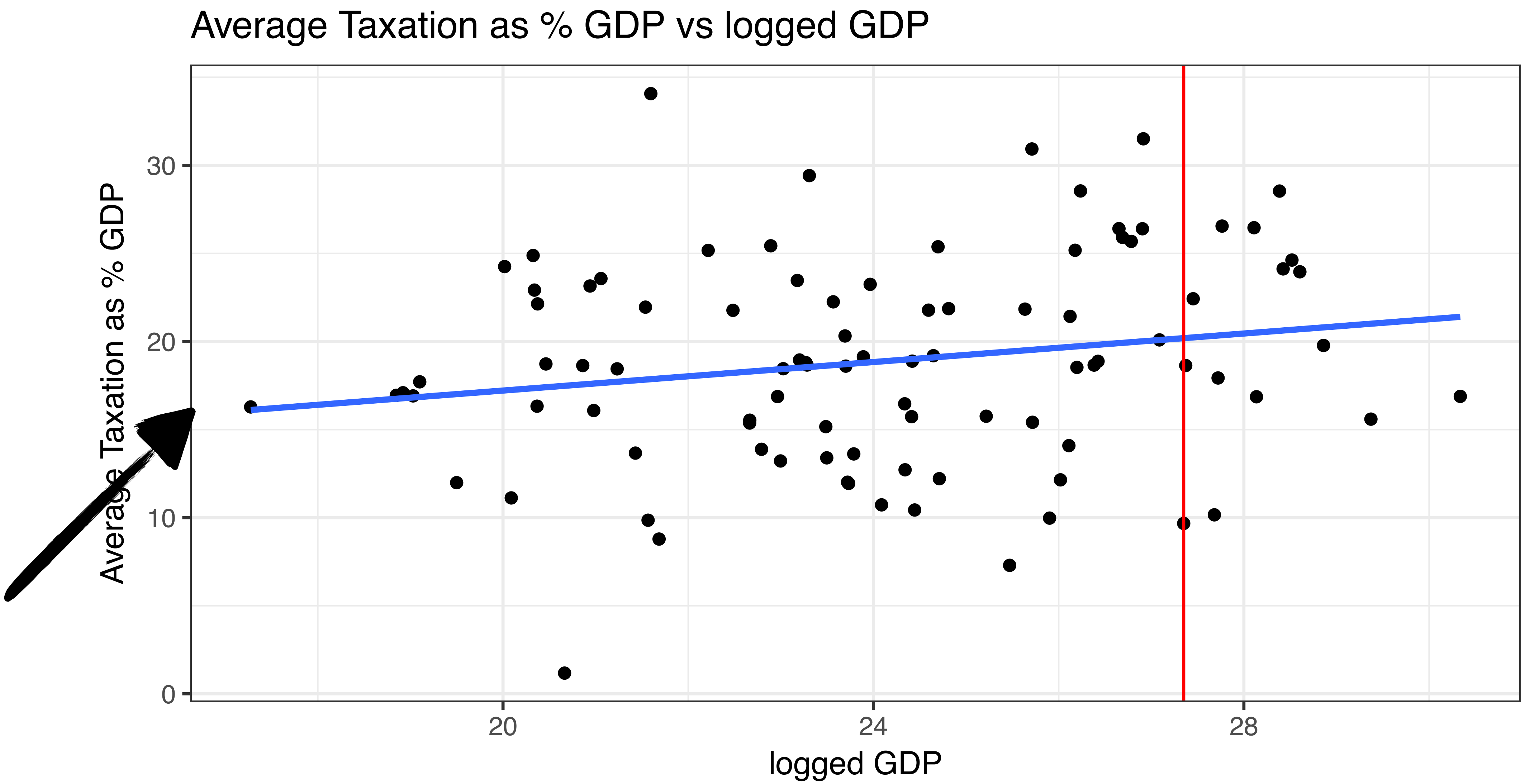


The full regression line

- Remember the formula of a line: $y = mx + b$
- So far we have only talked about m
- But what about b?

b (the intercept) is the point on y where the line crosses the x axis at zero

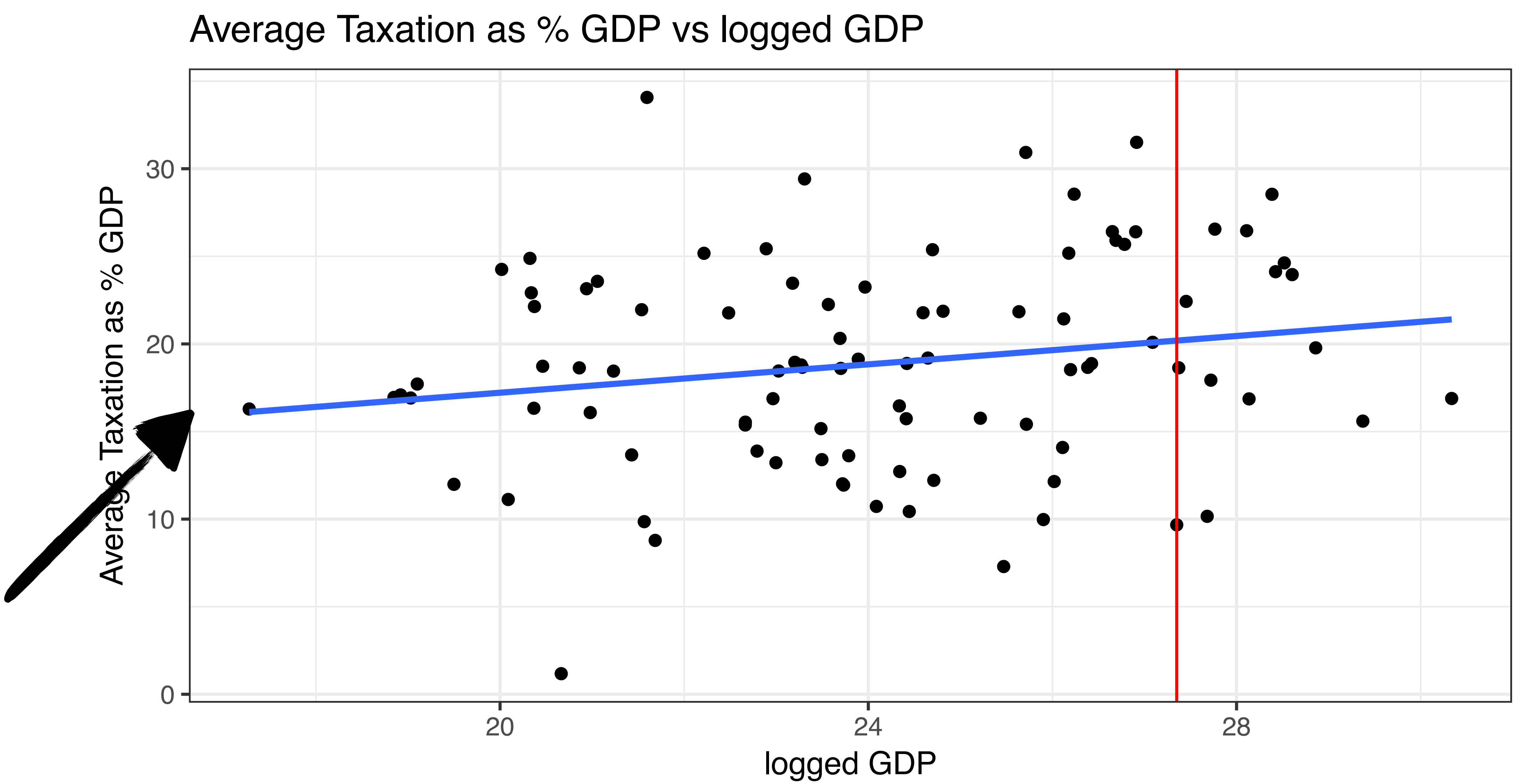




Finding the intercept

1. we find the slope

2. Then we find Y at x=0



Mean_y = 18.87, SD_y = 5.92

Mean_x = 24.09,SD_x = 2.84, r = 0.19

- The intercept does not always mean much
- It might be outside of the range of reasonable cases
- For example, predicting weight from height, a height of 0 makes little sense

Multiple regression

- Often we have additional variables that should be used in our model
- There might be things that are confounding factors for the relationship we are interested in
- The regression actually allows us to add other variables and "control" for these confounders

Multiple regression

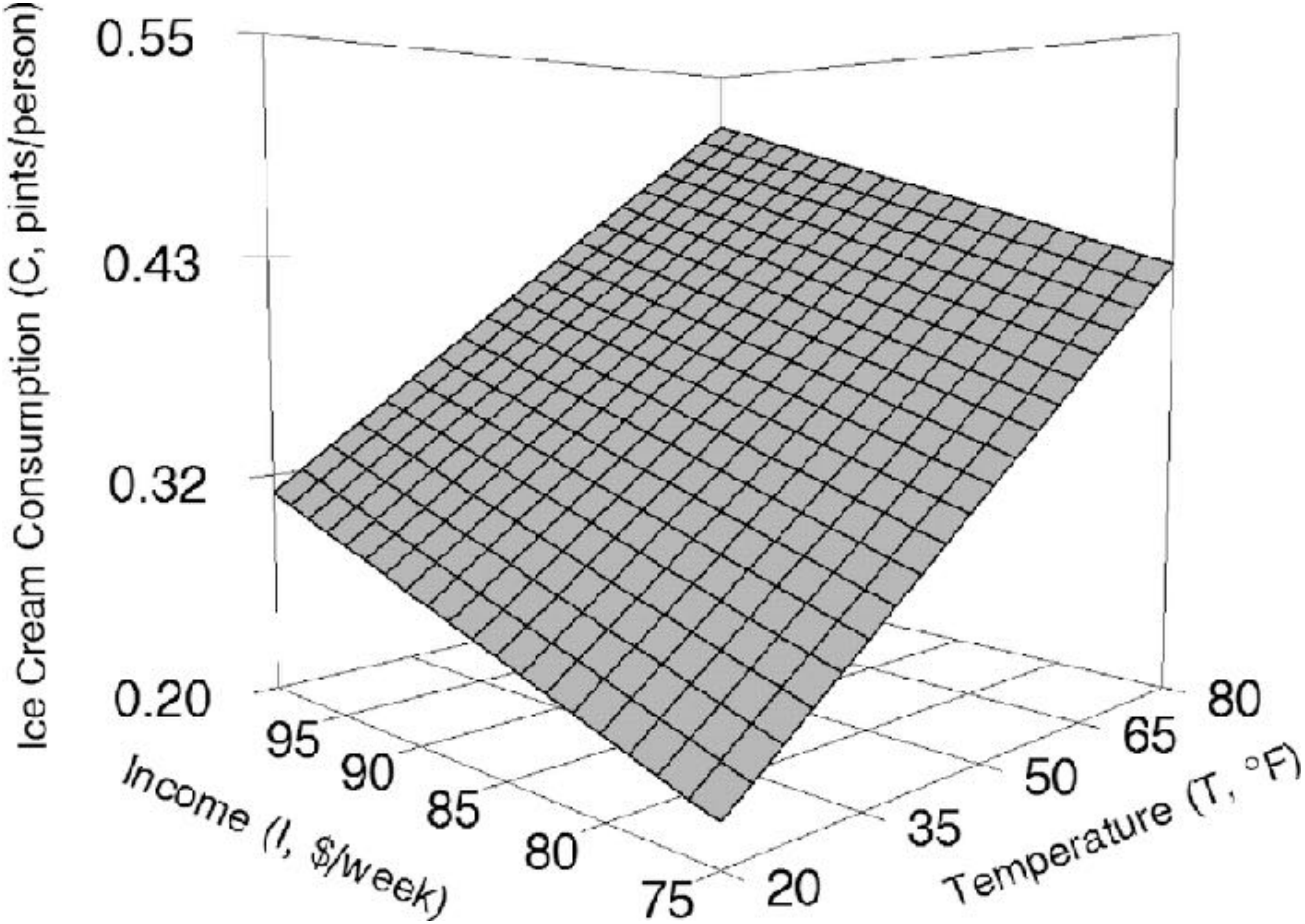
- let’s say we estimate effect of income on voting
  - But education level might matter too!
  - We can include both in the regression model!
- Estimate effect of smoking on life expectancy
  - Might want to control for exercise, nutrition, family health

Multiple regression

- In two-variable case line was drawn to minimize the error for each point

- Multiple regression is the same, but we are in a higher dimensional space (!!)

Multiple regression



Notation/Interpretation

$$Y = \alpha + \beta X$$
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots$$

- Each coefficient (beta) in the multiple regression is the linear change associated with a change of 1 in the associated variable, but holding all other variables constant

- Alpha is the intercept, or the predicted value when all X are equal to zero