PSC 400
SYRACUSE UNIVERSITY

DATA ANALYTICS FOR POLITICAL SCIENCE

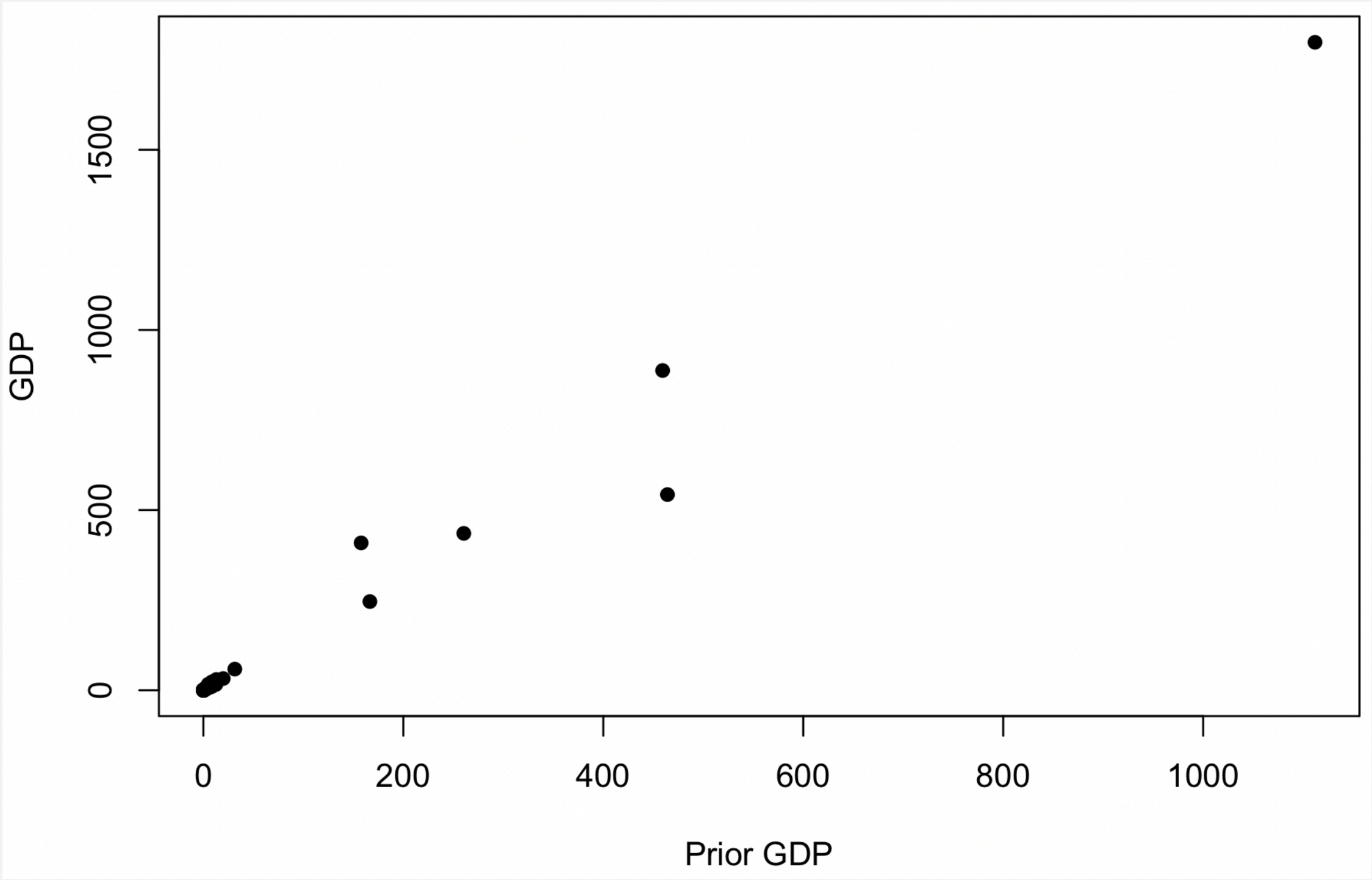
LINEAR REGRESSION

GDP

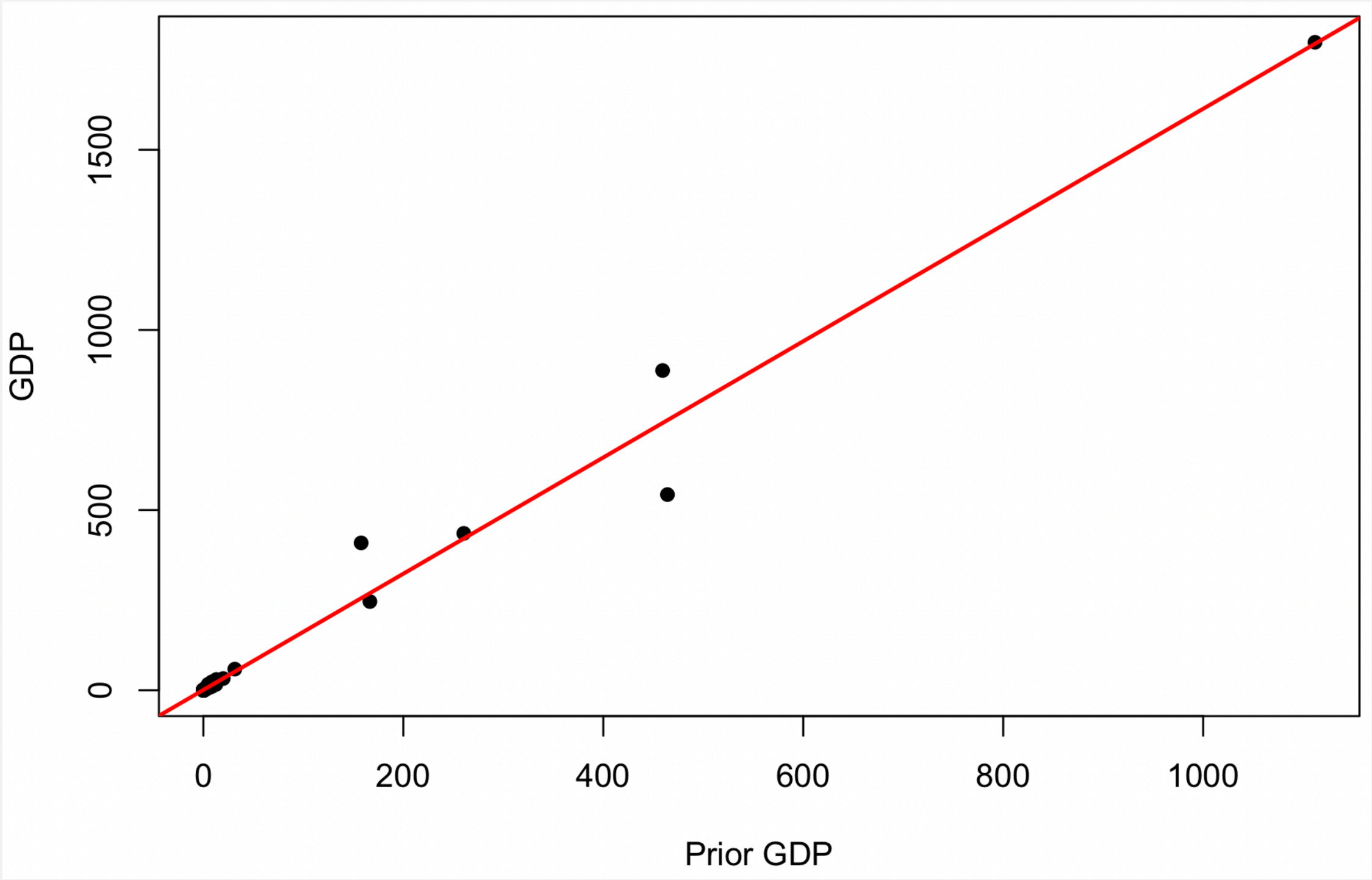
variable	description
country	name of the country
gdp	country's GDP from 2005 to 2006 (in trillions of local currency units)
prior_gdp	country's GDP from 1992 to 1993 (in trillions of local currency units)
light	country's average level of night-time light emissions from 2005 to 2006 (in units on a scale from 0 to 63, where 0 is complete darkness and 63 is extremely bright light)
prior_light	country's average level of night-time light emissions from 1992 to 1993 (in units on a scale from 0 to 63, where 0 is complete darkness and 63 is extremely bright light)

- countries.csv
- Create a scatterplot of prior gdp (x-axis) and gdp (y-axis)

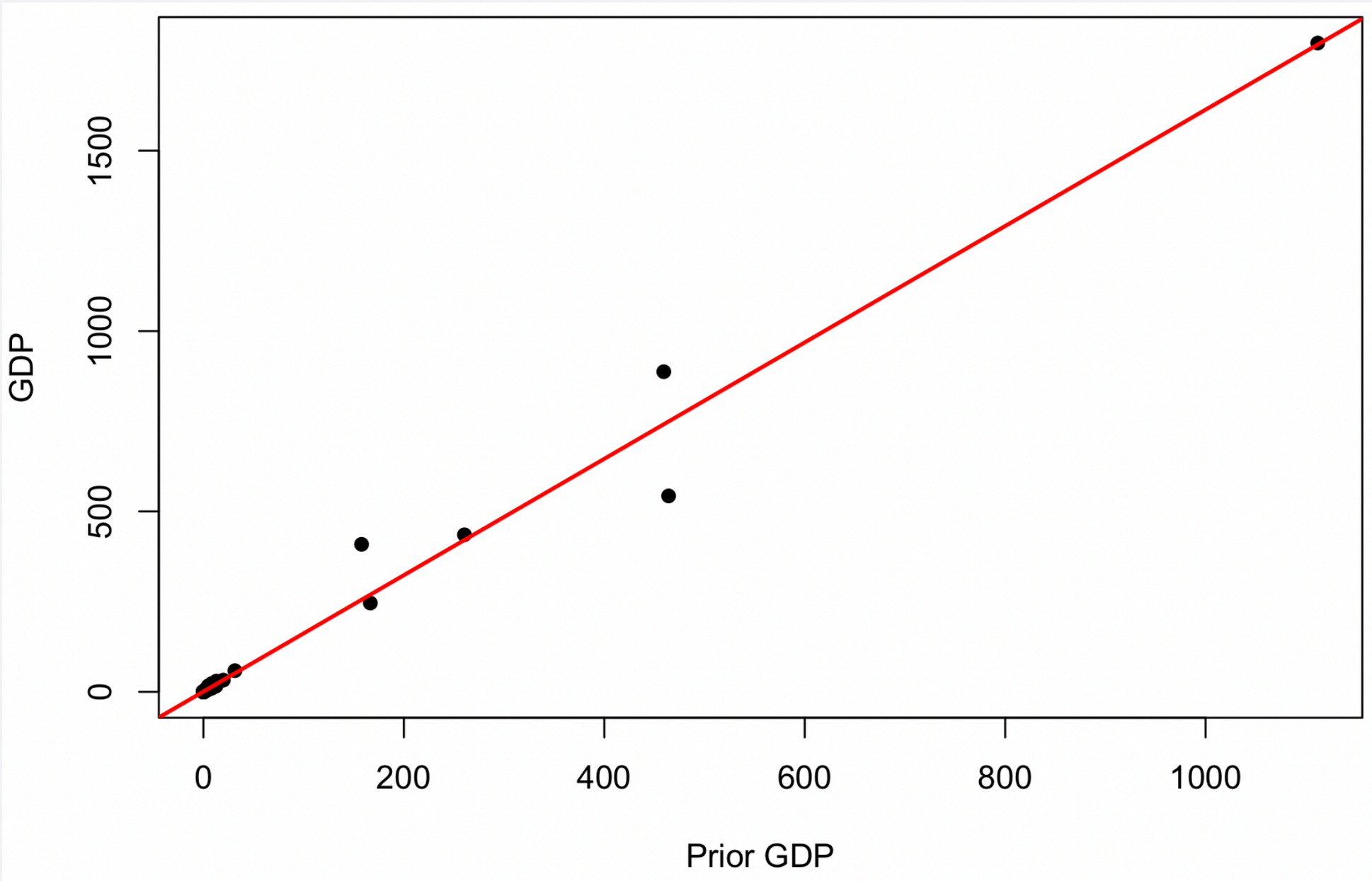
GDP



GDP

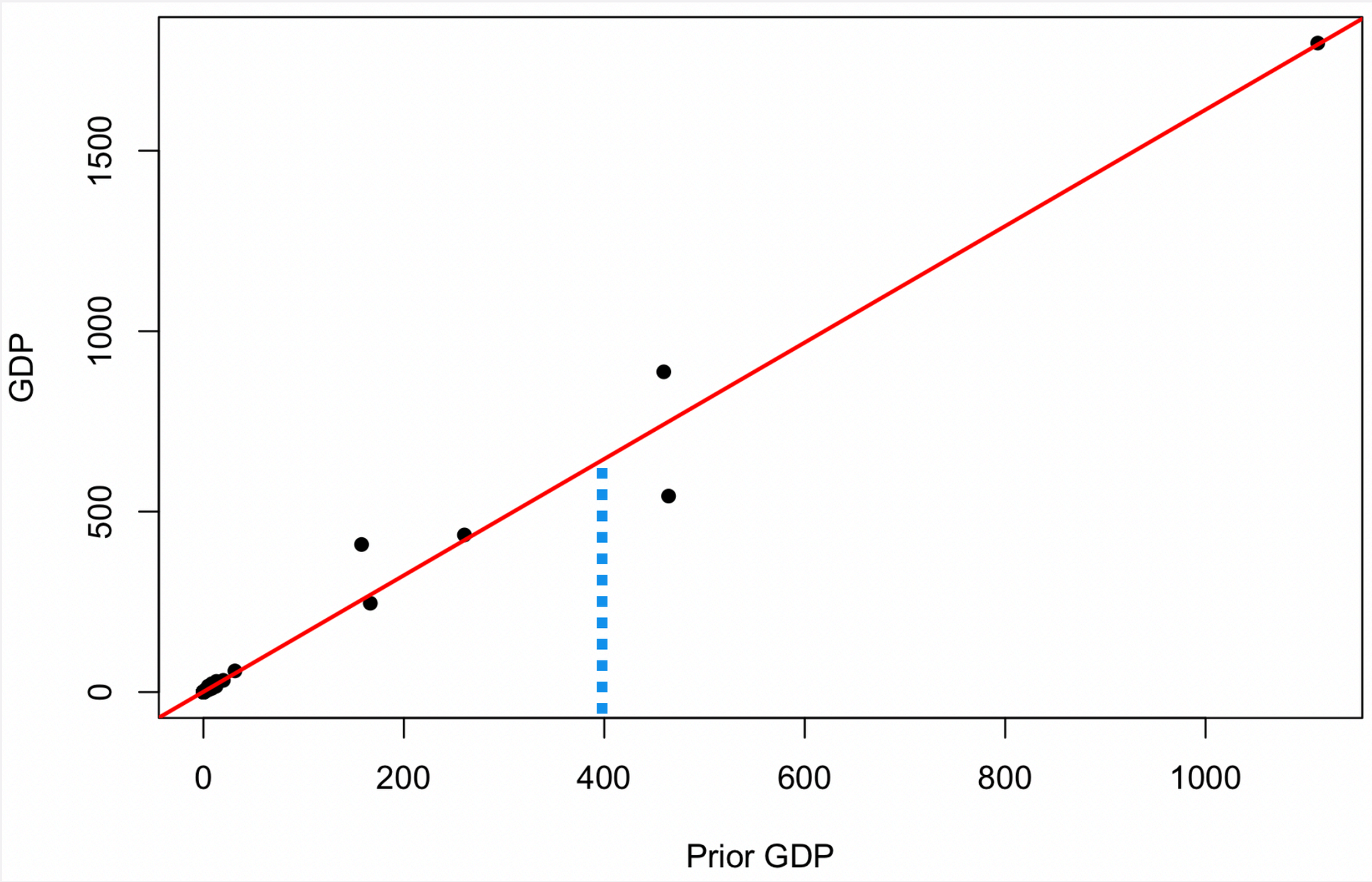


GDP



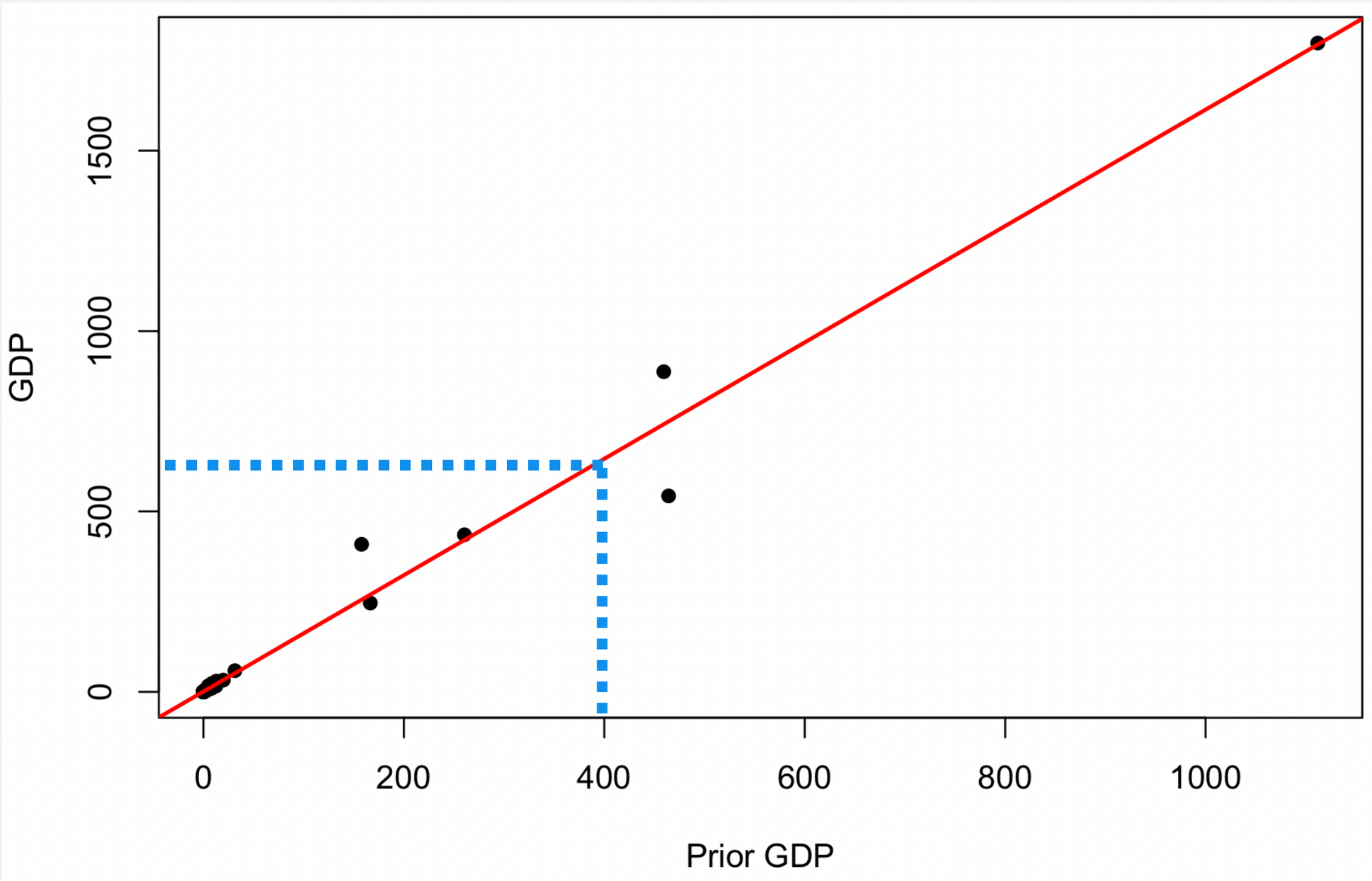
- **On average, how much higher is the current GDP of a country whose prior GDP was 400 instead of 600?**

GDP



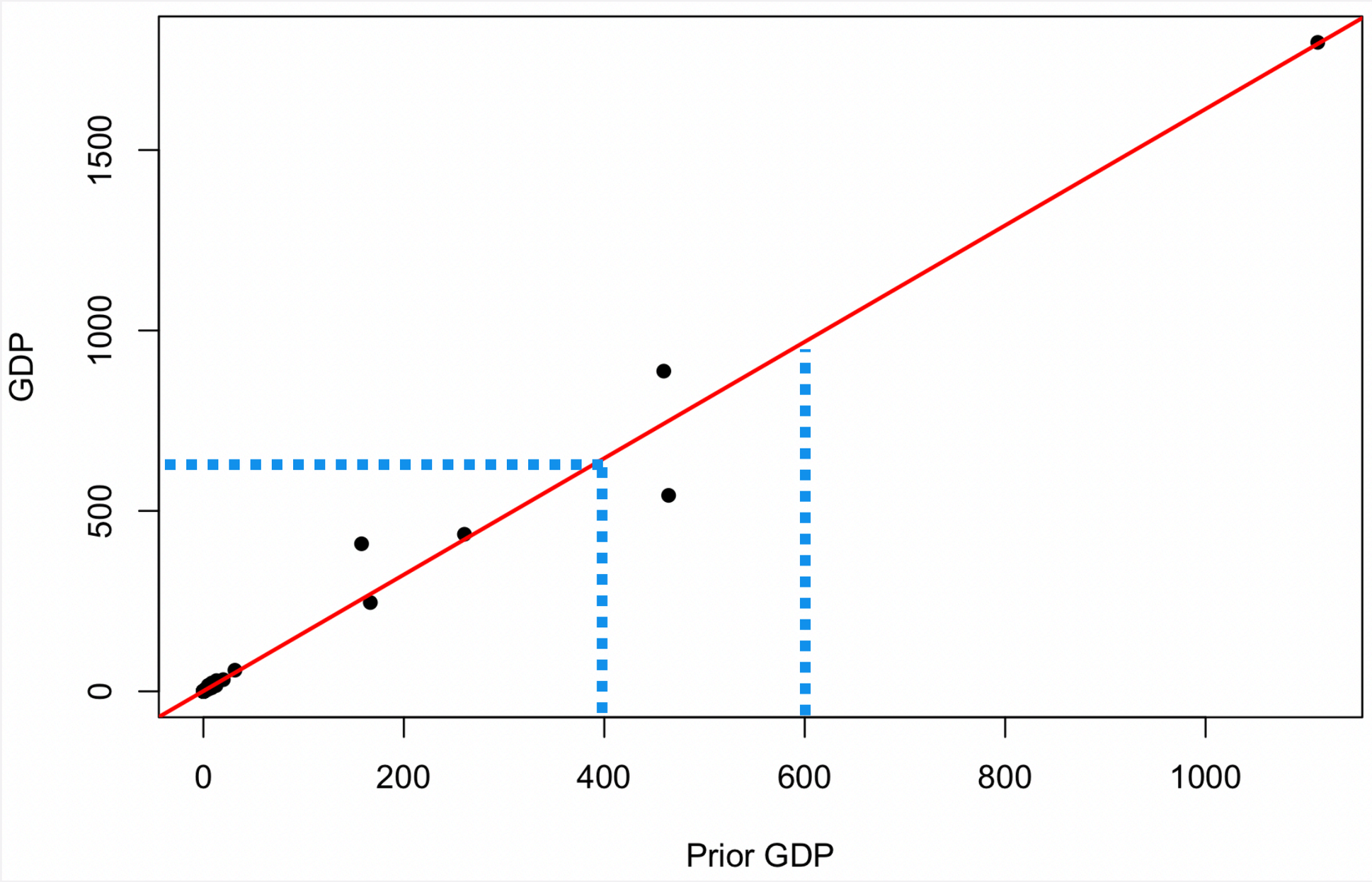
- **On average, how much higher is the current GDP of a country whose prior GDP was 400 instead of 600?**

GDP



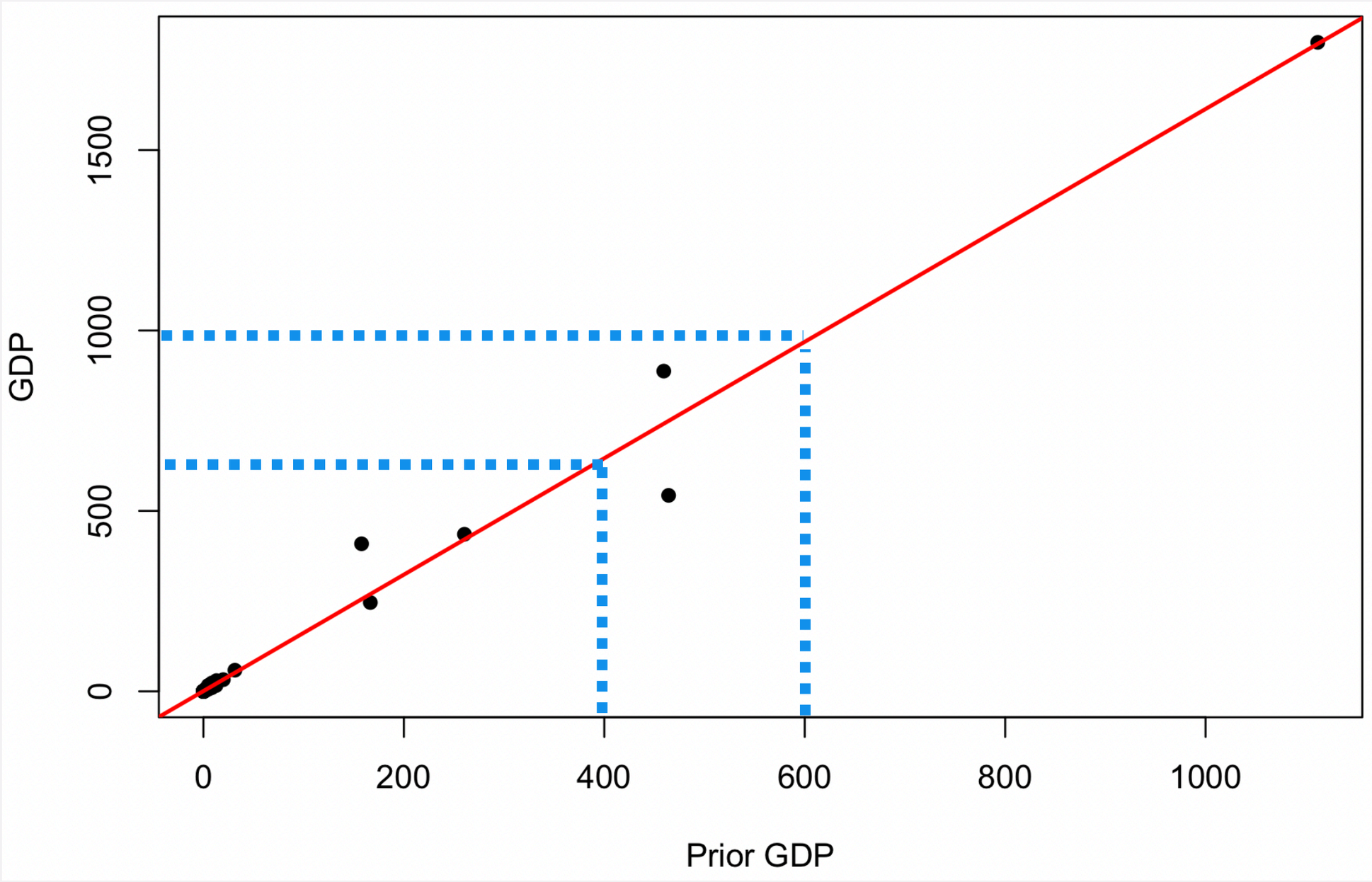
- **On average, how much higher is the current GDP of a country whose prior GDP was 400 instead of 600?**

GDP



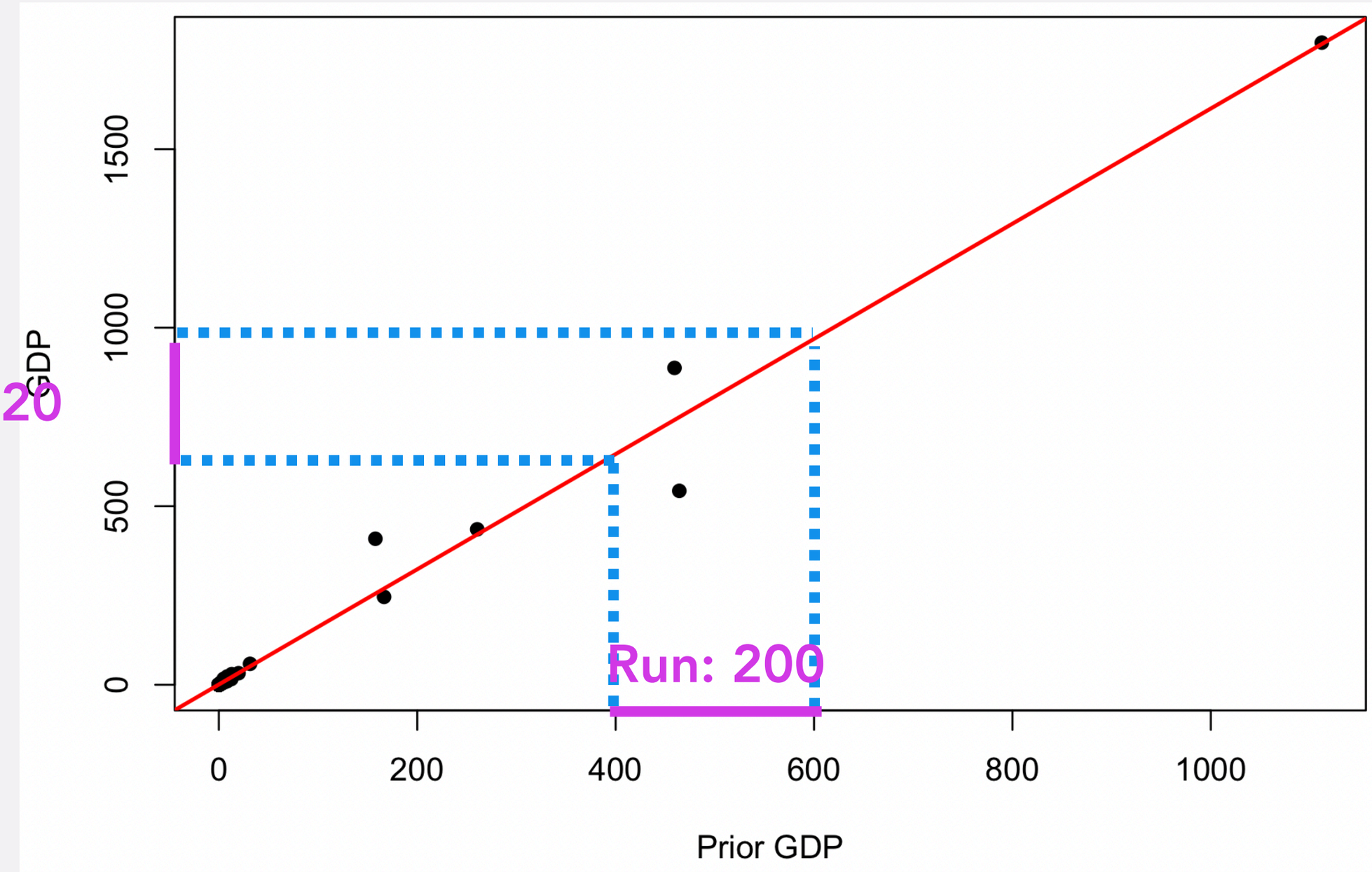
- **On average, how much higher is the current GDP of a country whose prior GDP was 400 instead of 600?**

GDP



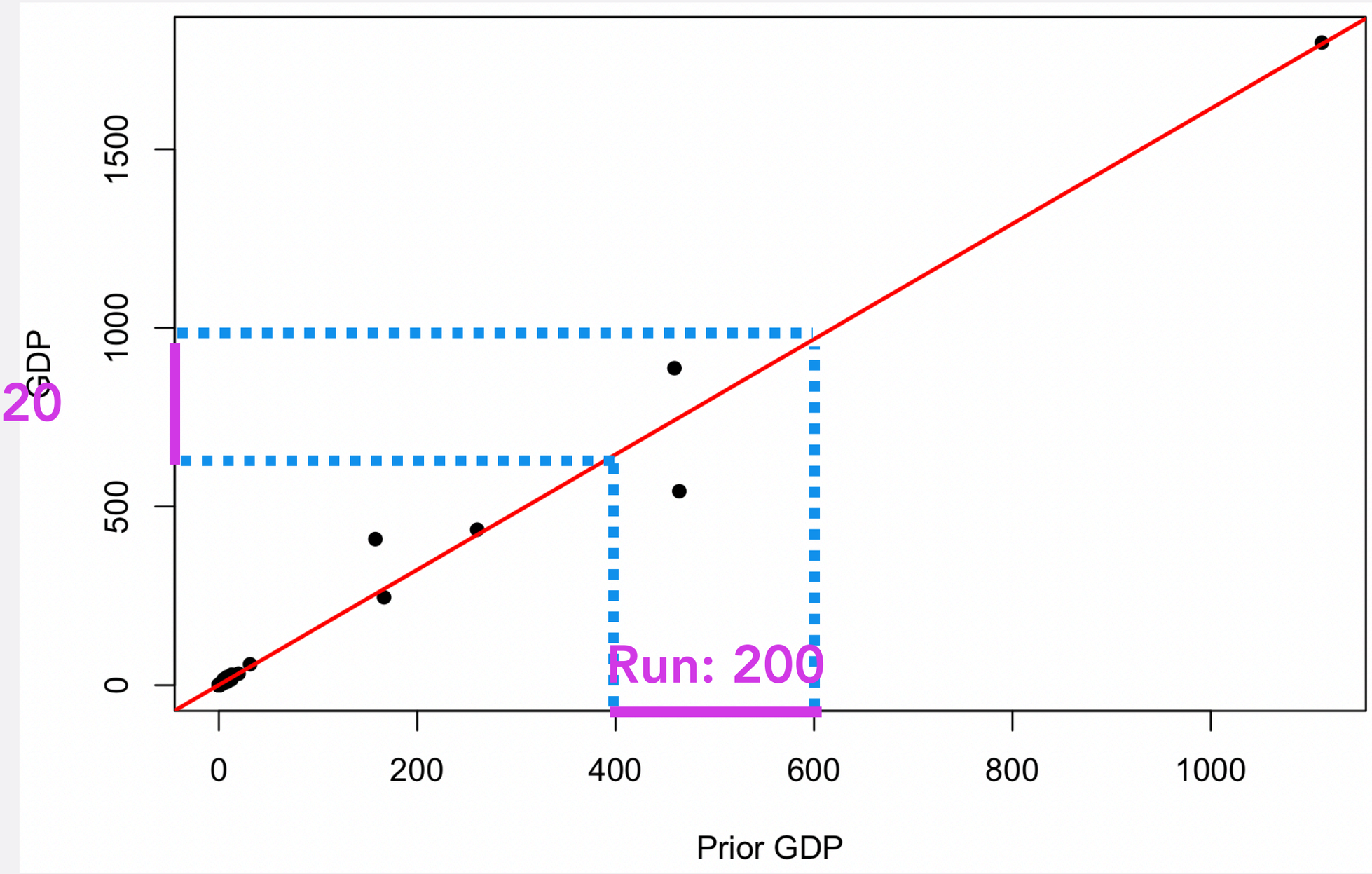
- **On average, how much higher is the current GDP of a country whose prior GDP was 400 instead of 600?**

GDP



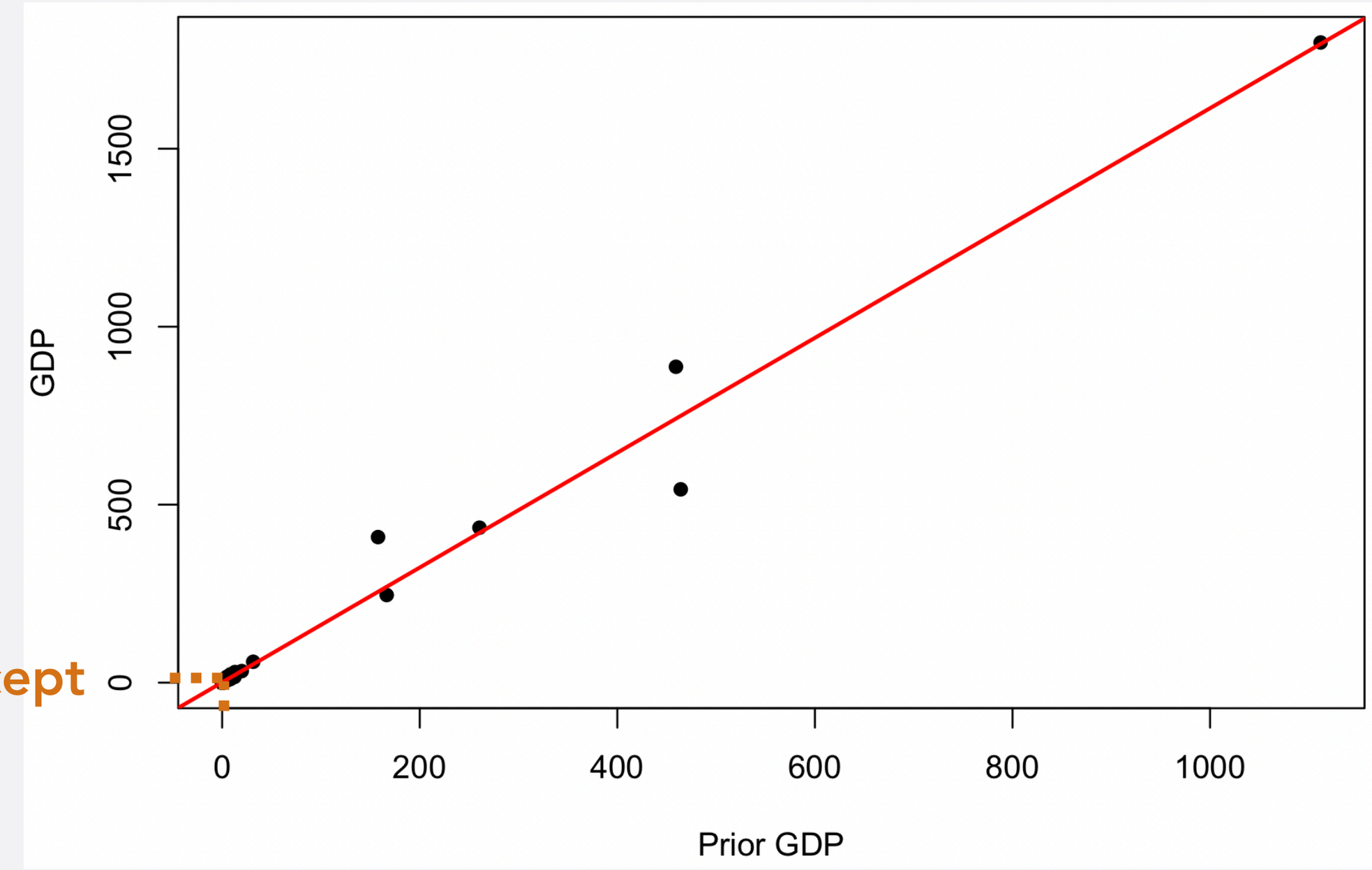
Slope=Rise over run=320/200=1.6

GDP



- For every one unit increase in prior GDP, current GDP is expected to increase by 1.6 (320/200)

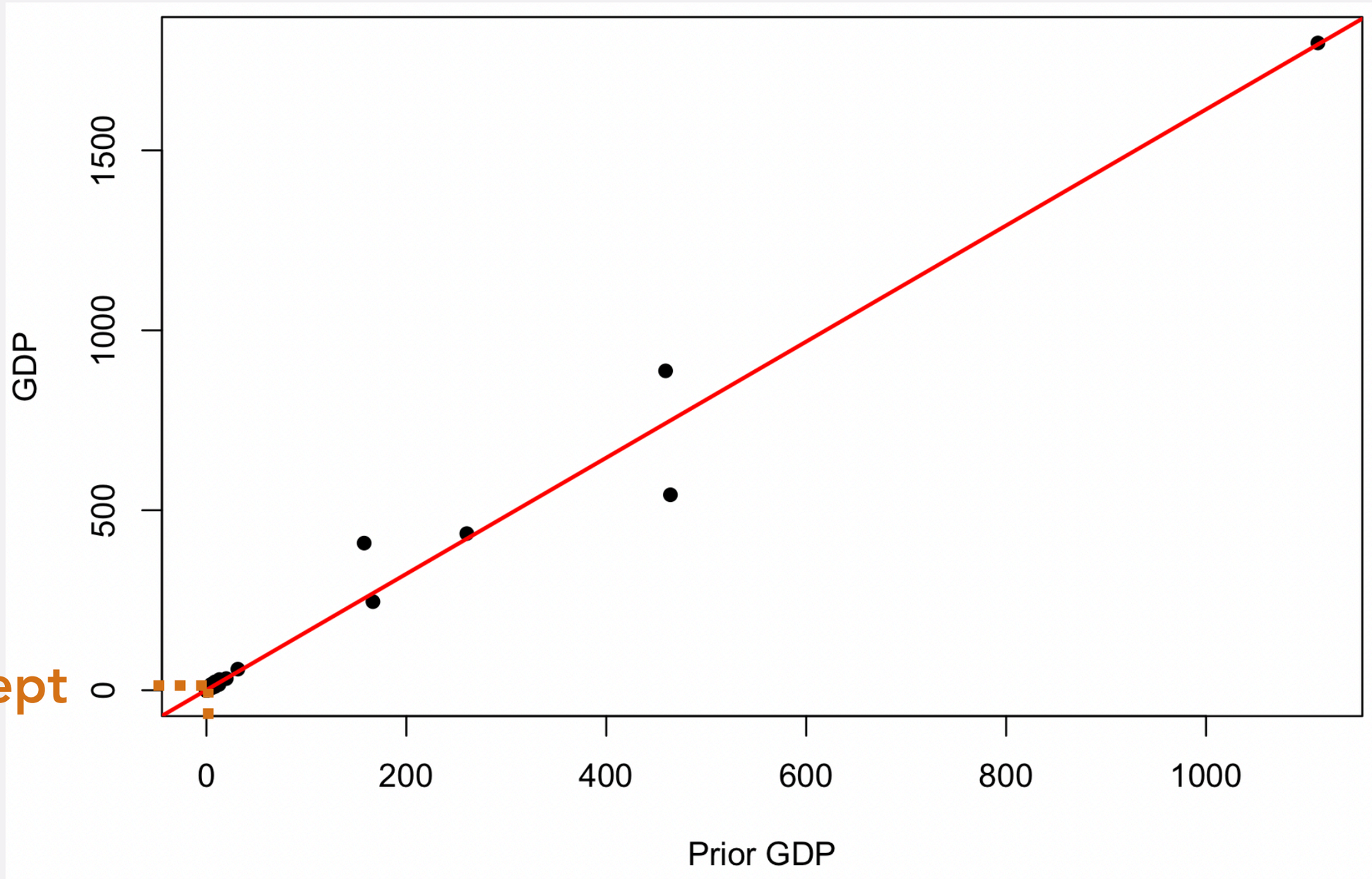
GDP



Intercept

Intercept=0.7

GDP



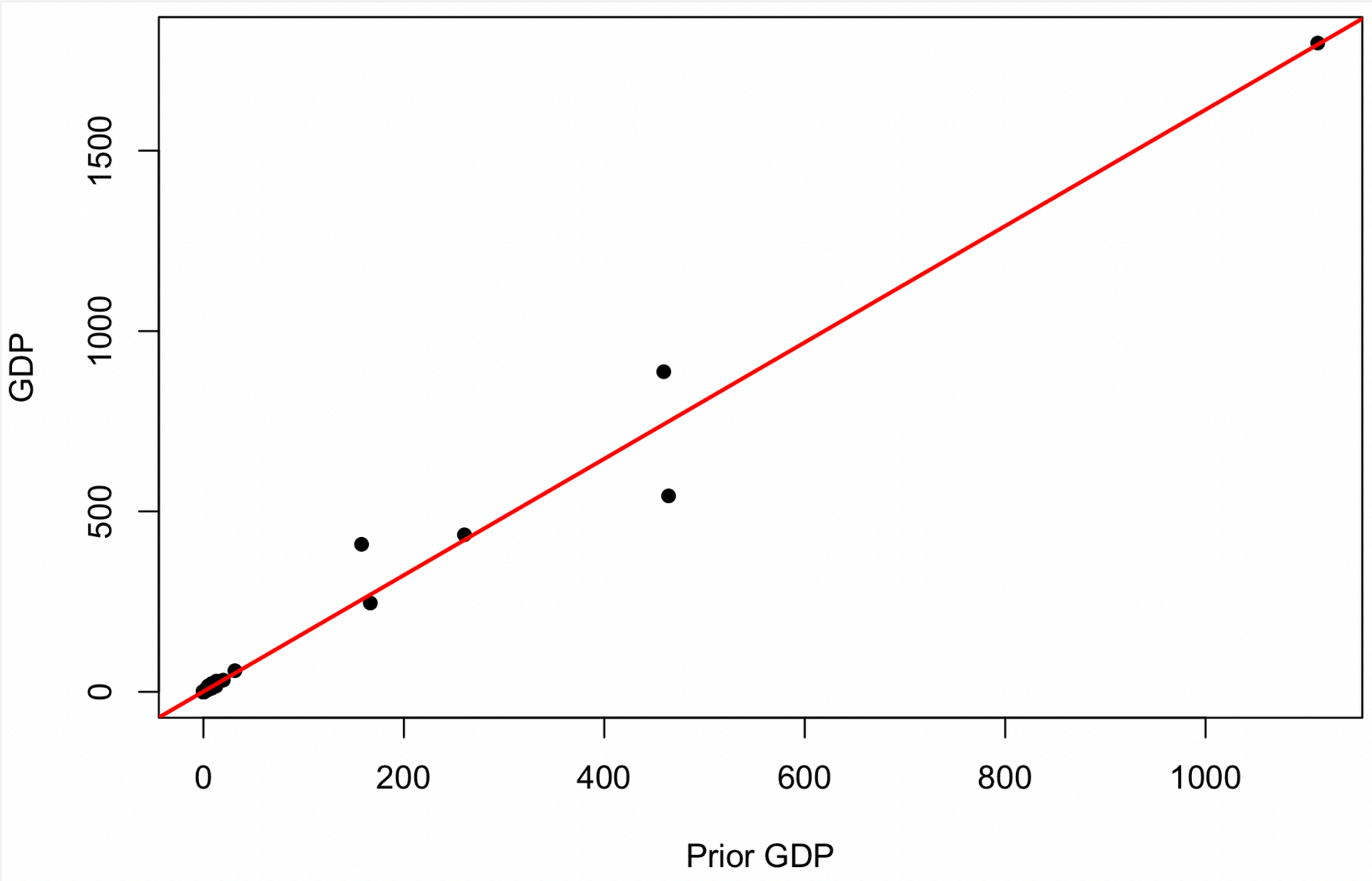
Intercept

- If prior GDP is 0, GDP is expected to be 0.7

LINEAR REGRESSION

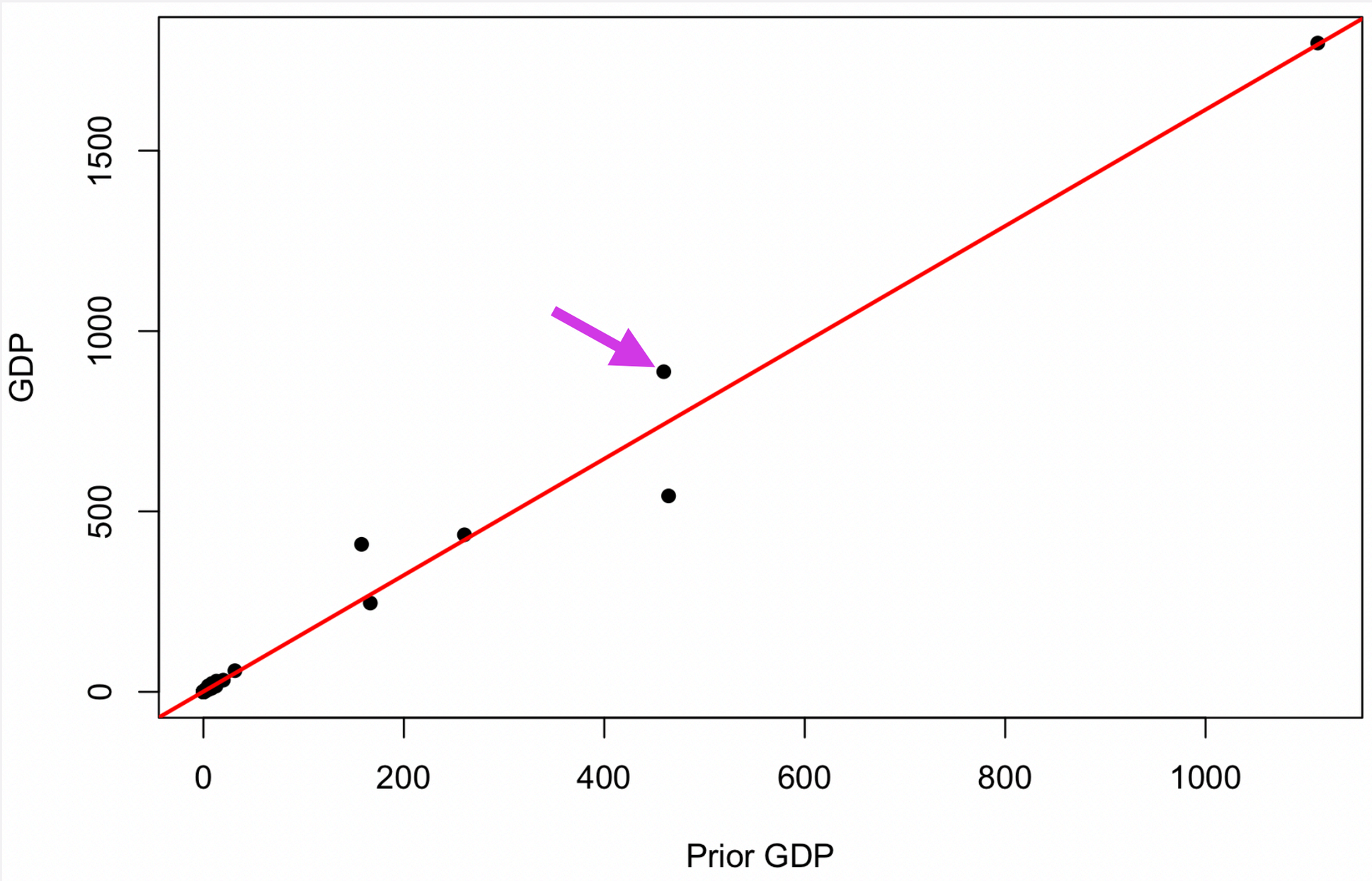
- Linear regression: Equation that tells us *direction* and *size* of relationship between independent variable (IV) and dependent variable (DV)
- DV = Intercept + Slope * IV + error

GDP



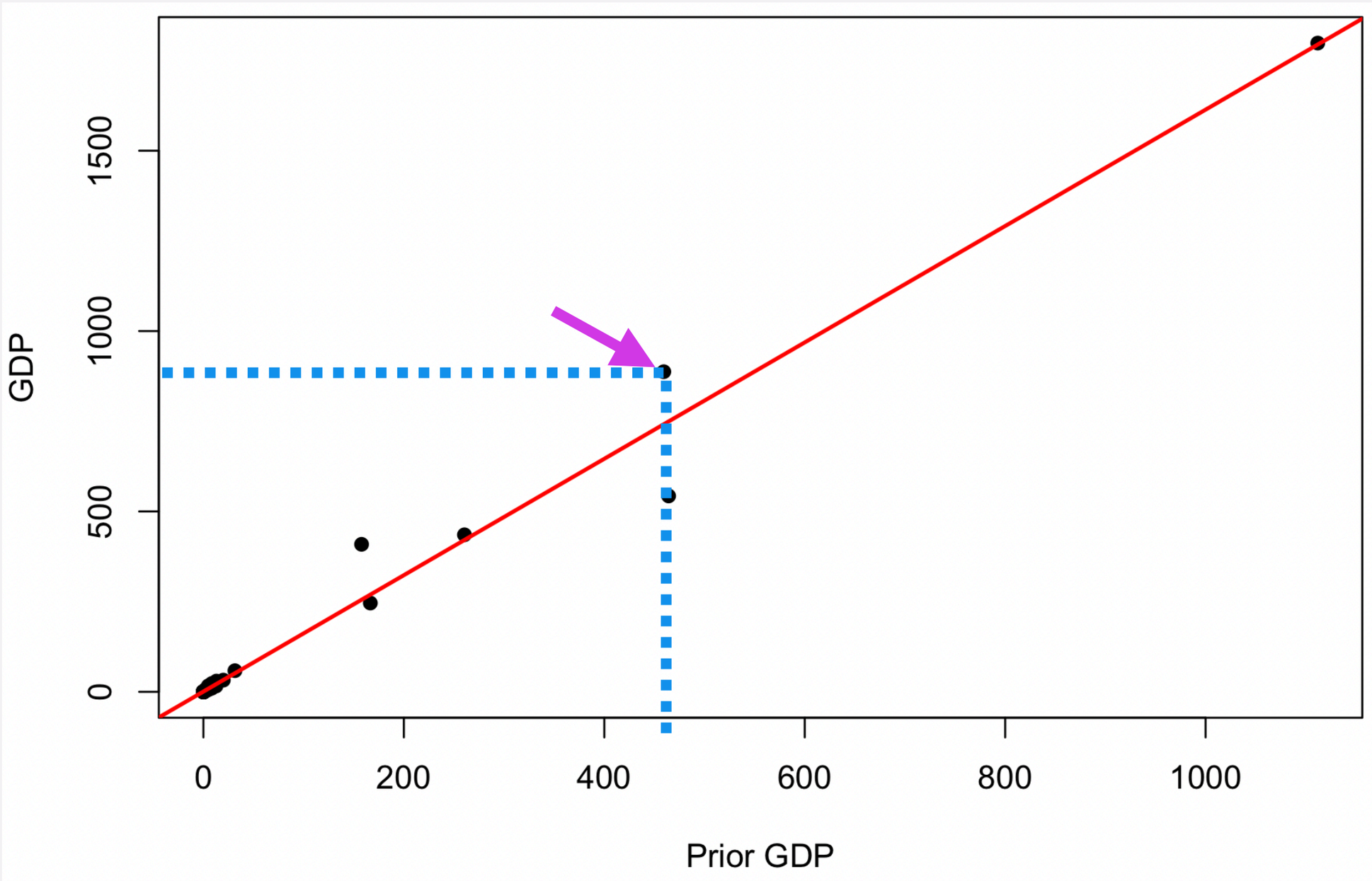
- GDP = 0.7 + 1.6 * Prior GDP + error

GDP



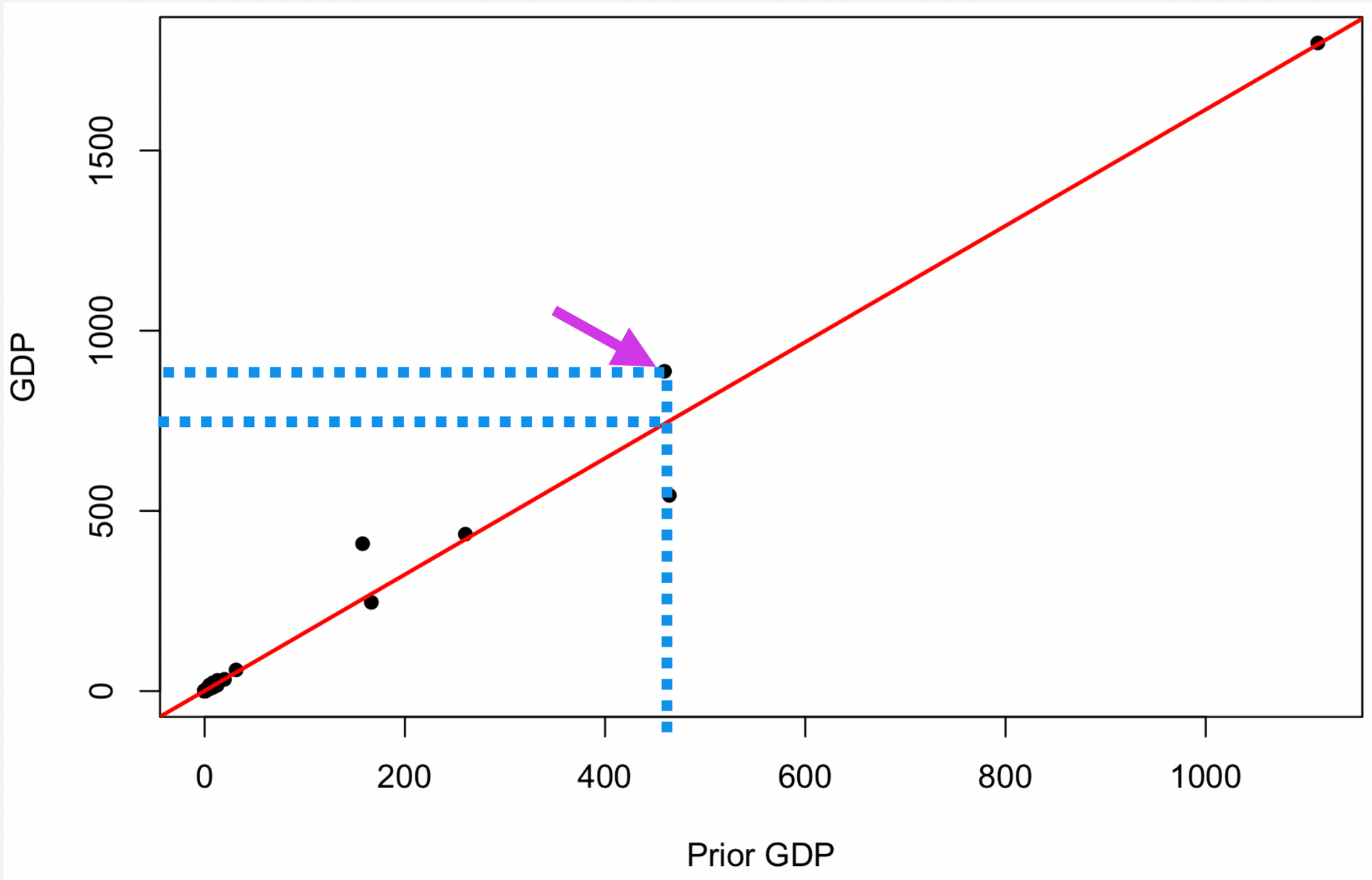
- GDP = 0.7 + 1.6 * Prior GDP + error

GDP



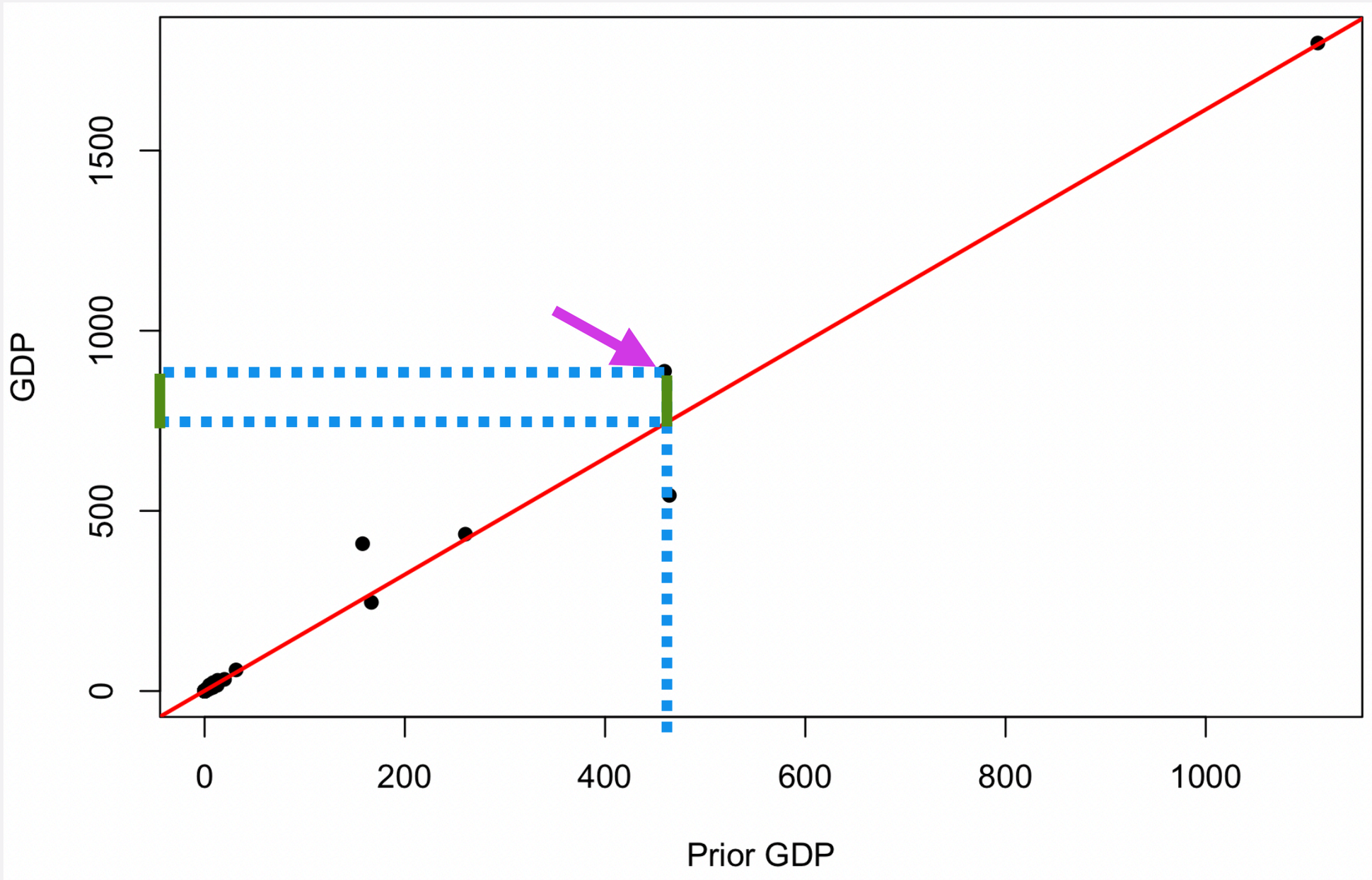
- GDP = 0.7 + 1.6 * Prior GDP + error

GDP



- GDP = 0.7 + 1.6 * Prior GDP + error

GDP



- GDP = 0.7 + 1.6 * Prior GDP + error

PREDICTION ERROR

- For each observation, we have a prediction error: $y - \hat{y}$
  - y: actual observed value
  - $\hat{y}$: predicted value (by regressions line)
  - $y - \hat{y}$: prediction error, residual
- We square the prediction errors: $(y - \hat{y})^2$
  - 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

BEST LINE

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

EXAMPLE

Table 4.5. 2012 US Presidential Election Data.

Variable	*Description*
`state`	abbreviated name of the state
`Obama`	Obama's vote share (percentage)
`Romney`	Romney's vote share (percentage)
`EV`	number of Electoral College votes for the state

- pres12.csv

- How does Obama's vote share in 2012 depend on his 2008 vote share?

EXAMPLE

Table 4.1. 2008 US Presidential Election Data.

Variable	*Description*
`state`	abbreviated name of the state
`state.name`	unabbreviated name of the state
`Obama`	Obama's vote share (percentage)
`McCain`	McCain's vote share (percentage)
`EV`	number of Electoral College votes for the state

Table 4.5. 2012 US Presidential Election Data.

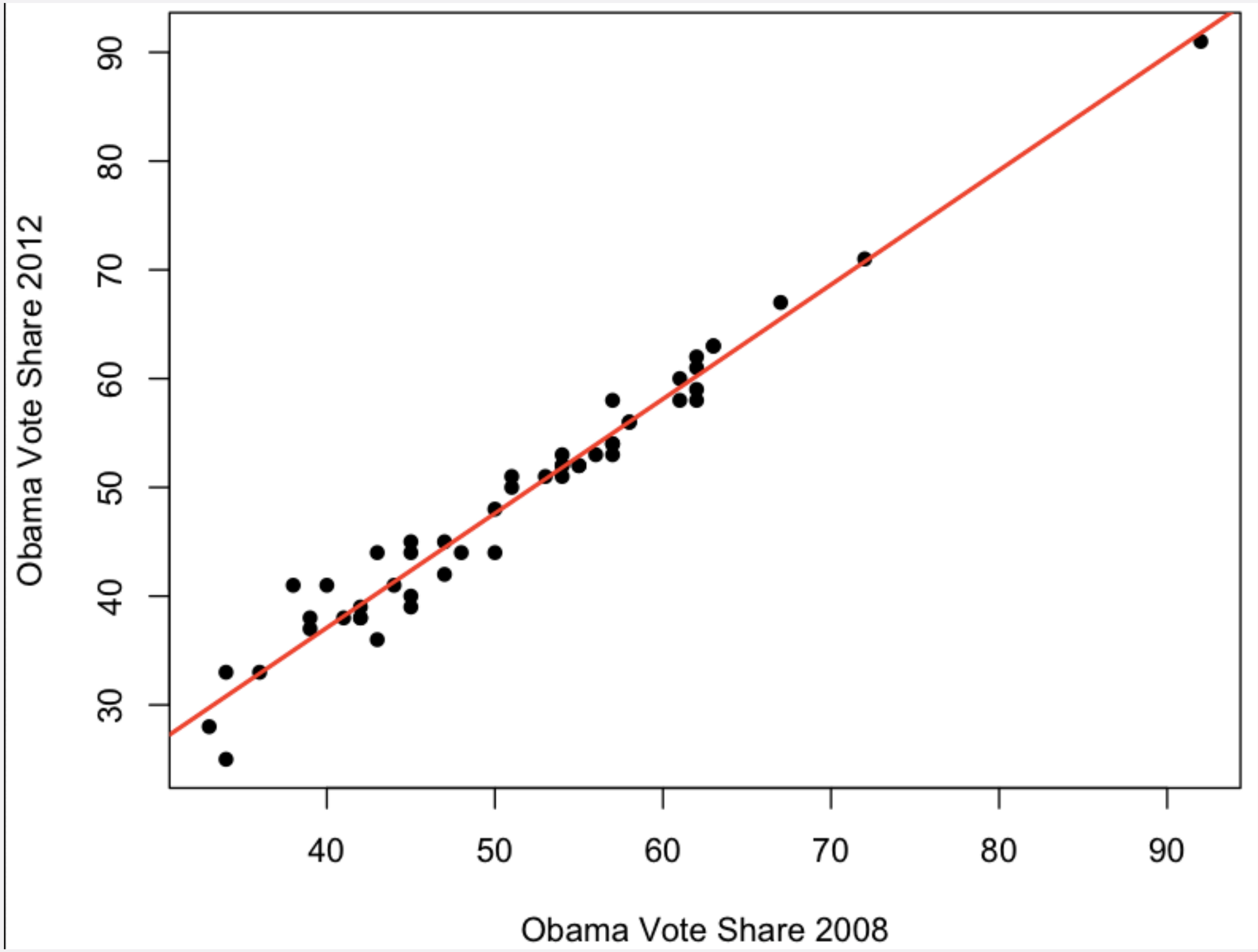
Variable	*Description*
`state`	abbreviated name of the state
`Obama`	Obama's vote share (percentage)
`Romney`	Romney's vote share (percentage)
`EV`	number of Electoral College votes for the state

- pres08.csv, pres12.csv

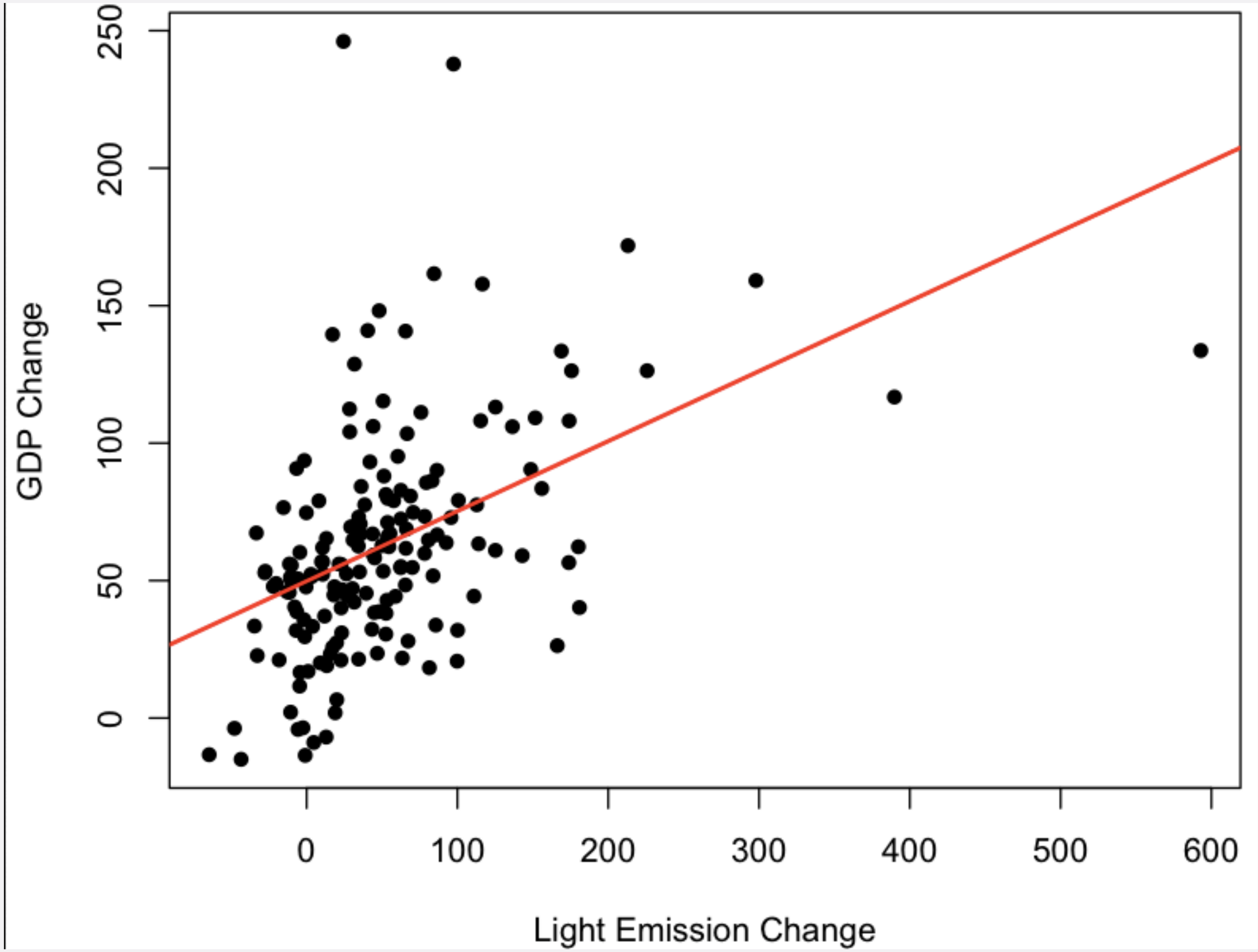
EXERCISE

- Create a scatterplot of Obama's 2008 vote share (x-axis) and 2012 vote share (y-axis)
- Estimate a linear regression to predict his 2012 vote share using his 2008 vote share
- Add the regression line to your scatterplot
- Interpret the intercept and slope coefficients
- What's the predicted 2012 vote share for a state where he got 50% in 2008?

COMPARE


Obama Vote Share 2012
30
40
50
60
70
80
90
40
50
60
70
80
90
Obama Vote Share 2008

COMPARE


GDP Change
250
200
150
100
50
0
0
100
200
300
400
500
600
Light Emission Change

EXPLANATORY POWER MEASURE

- **Need: measure of how well independent variable explains dependent variable in a linear regression**

EXPLANATORY POWER MEASURE

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