PSC 400
SYRACUSE UNIVERSITY

DATA ANALYTICS FOR POLITICAL SCIENCE

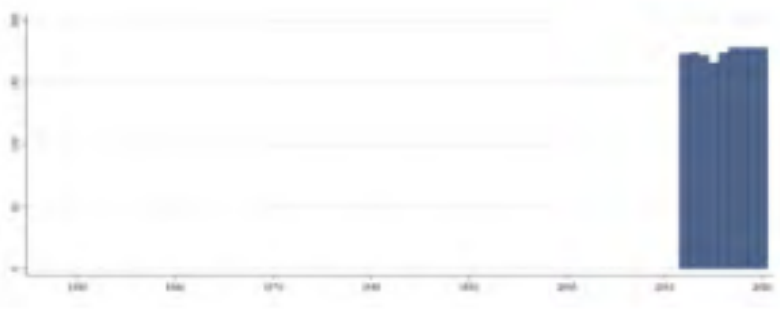
FINDING AND CLEANING DATA, LINEAR REGRESSION

EXERCISE: QOG

4.94.1 Corruption Perceptions Index (ti_cpi)

Corruption Perceptions Index. Scale of 0-100 where a 0 equals the highest level of perceived corruption and 100 equals the lowest level of perceived corruption.

Min. Year:2017 **Max. Year**: 2017
N: 178



Min. Year:2012 **Max. Year**: 2020
N: 178 **n**: 1571 $\overline{N}$: 175 $\overline{T}$: 9

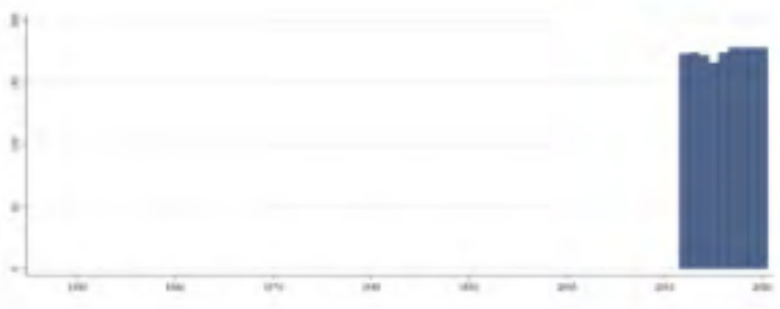
- Create a new variable ti_cpi_max10 where 0 is the highest level of corruption and 10 is the lowest

EXERCISE: QOG

4.94.1 Corruption Perceptions Index (ti_cpi)

Corruption Perceptions Index. Scale of 0-100 where a 0 equals the highest level of perceived corruption and 100 equals the lowest level of perceived corruption.

Min. Year:2017 **Max. Year**: 2017
N: 178



Min. Year:2012 **Max. Year**: 2020
N: 178 **n**: 1571 $\overline{N}$: 175 $\overline{T}$: 9

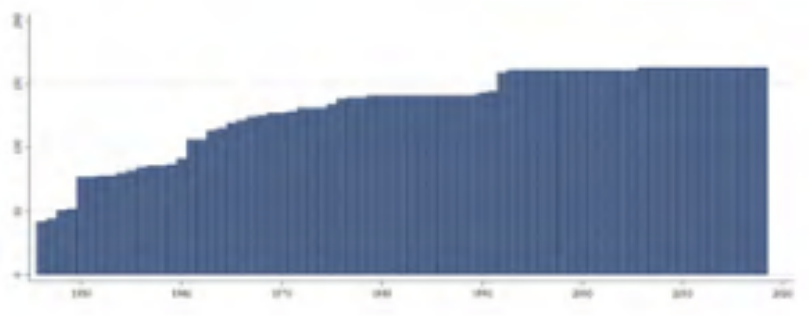
- Create a new variable ti_cpi_reverse where 100 is the *highest* level of corruption and 0 is the lowest

EXERCISE: QOG

4.70.1 Real GDP per Capita (mad_gdppc)

Real GDP per capita in 2011 US dollars, multiple benchmarks.

Min. Year:2017 **Max. Year**: 2017
N: 163



Min. Year:1946 **Max. Year**: 2018
N: 175 **n**: 9559 $\overline{N}$: 131 $\overline{T}$: 55

- Plot the density of mad_gdppc
- Create a new variable mad_gdppc_log that is the logged value of mad_gdppc (function: log)
- Plot the density of mad_gdppc_logrr

EXERCISE: QOG

4.41.1 Colonial Origin (ht_colonial)

This is a tenfold classification of the former colonial ruler of the country. Following Bernard et al. (2004), we have excluded the British settler colonies (the US, Canada, Australia, Israel and New Zealand), and exclusively focused on "Western overseas" colonialism. This implies that only Western colonizers (e.g. excluding Japanese colonialism), and only countries located in the non-Western hemisphere "overseas" (e.g. excluding Ireland & Malta), have been coded. Each country that has been colonized since 1700 is coded. In cases of several colonial powers, the last one is counted, if it lasted for 10 years or longer. The categories are the following:

0. Never colonized by a Western overseas colonial power
1. Dutch
2. Spanish
3. Italian
4. US
5. British
6. French
7. Portuguese
8. Belgian
9. British-French
10. Australian

EXERCISE: QOG

4.41.1 Colonial Origin (ht_colonial)

This is a tenfold classification of the former colonial ruler of the country. Following Bernard et al. (2004), we have excluded the British settler colonies (the US, Canada, Australia, Israel and New Zealand), and exclusively focused on "Western overseas" colonialism. This implies that only Western colonizers (e.g. excluding Japanese colonialism), and only countries located in the non-Western hemisphere "overseas" (e.g. excluding Ireland & Malta), have been coded. Each country that has been colonized since 1700 is coded. In cases of several colonial powers, the last one is counted, if it lasted for 10 years or longer. The categories are the following:

0. Never colonized by a Western overseas colonial power
1. Dutch
2. Spanish
3. Italian
4. US
5. British
6. French
7. Portuguese
8. Belgian
9. British-French
10. Australian

- Create a variable "colonized" that is 1 if the country was every colonized by a Western colonial power, and 0 if not

EXERCISE: QOG

4.41.1 Colonial Origin (ht_colonial)

This is a tenfold classification of the former colonial ruler of the country. Following Bernard et al. (2004), we have excluded the British settler colonies (the US, Canada, Australia, Israel and New Zealand), and exclusively focused on "Western overseas" colonialism. This implies that only Western colonizers (e.g. excluding Japanese colonialism), and only countries located in the non-Western hemisphere "overseas" (e.g. excluding Ireland & Malta), have been coded. Each country that has been colonized since 1700 is coded. In cases of several colonial powers, the last one is counted, if it lasted for 10 years or longer. The categories are the following:

0. Never colonized by a Western overseas colonial power
1. Dutch
2. Spanish
3. Italian
4. US
5. British
6. French
7. Portuguese
8. Belgian
9. British-French
10. Australian

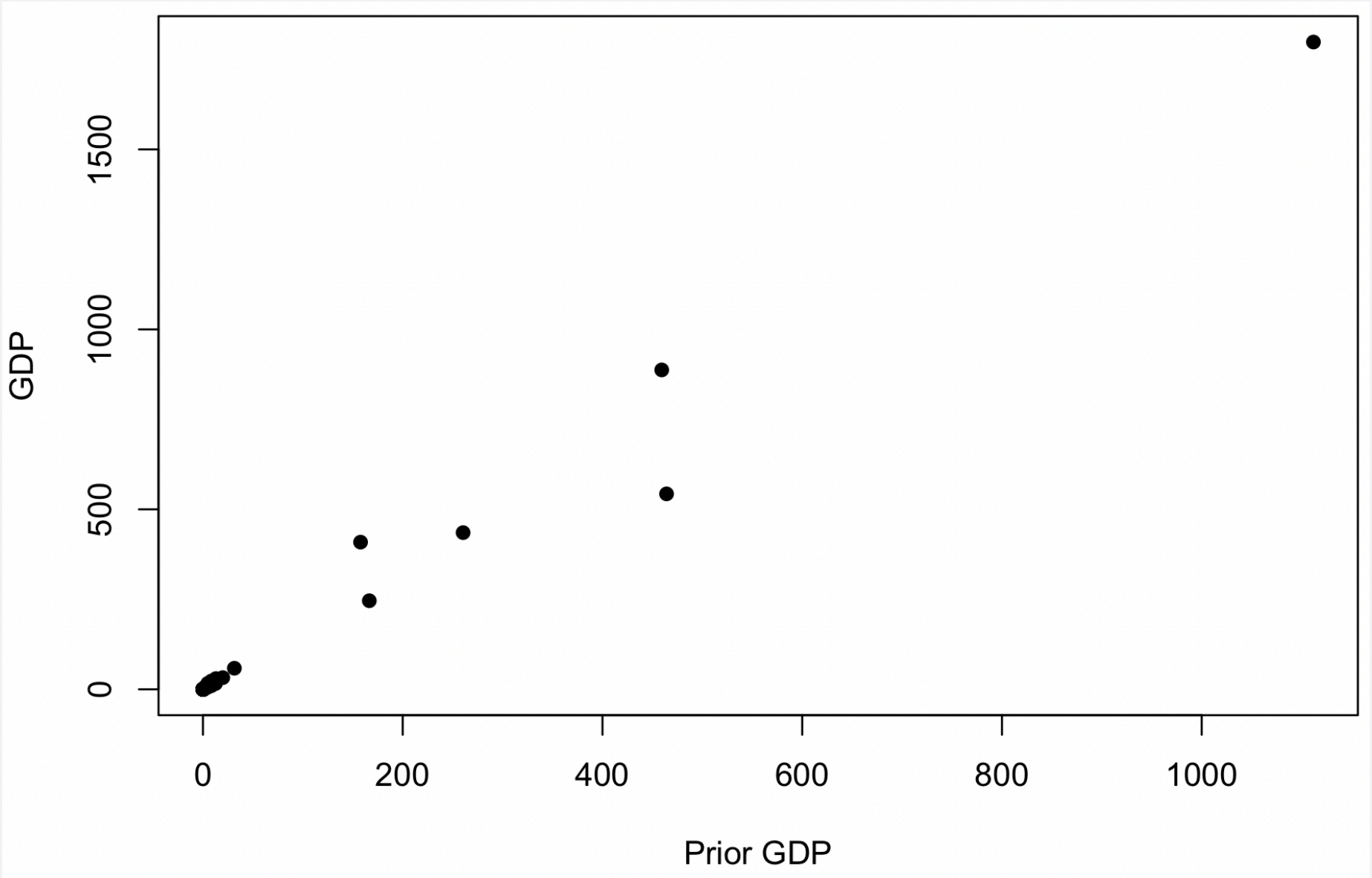
- Create a variable "colonized2" that is "colonized" if the country was ever colonized by a Western colonial power, and "not colonized" if not

GDP

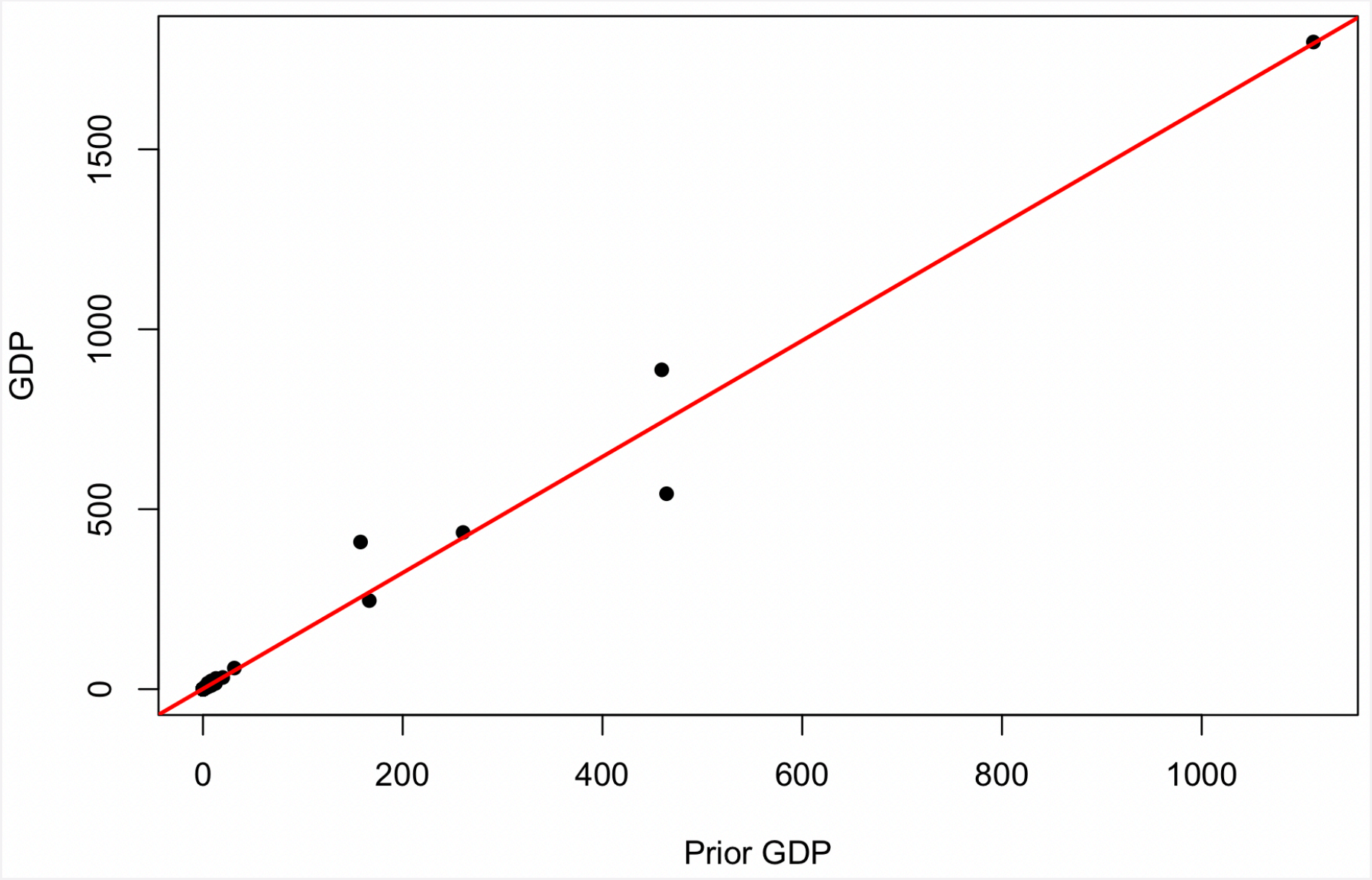
variable	description
country	name of the country
gdp	country's GDP from 2005 to 2006 (in trillions of local currency units)
prior_gdp	country's GDP from 1992 to 1993 (in trillions of local currency units)
light	country's average level of night-time light emissions from 2005 to 2006 (in units on a scale 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)

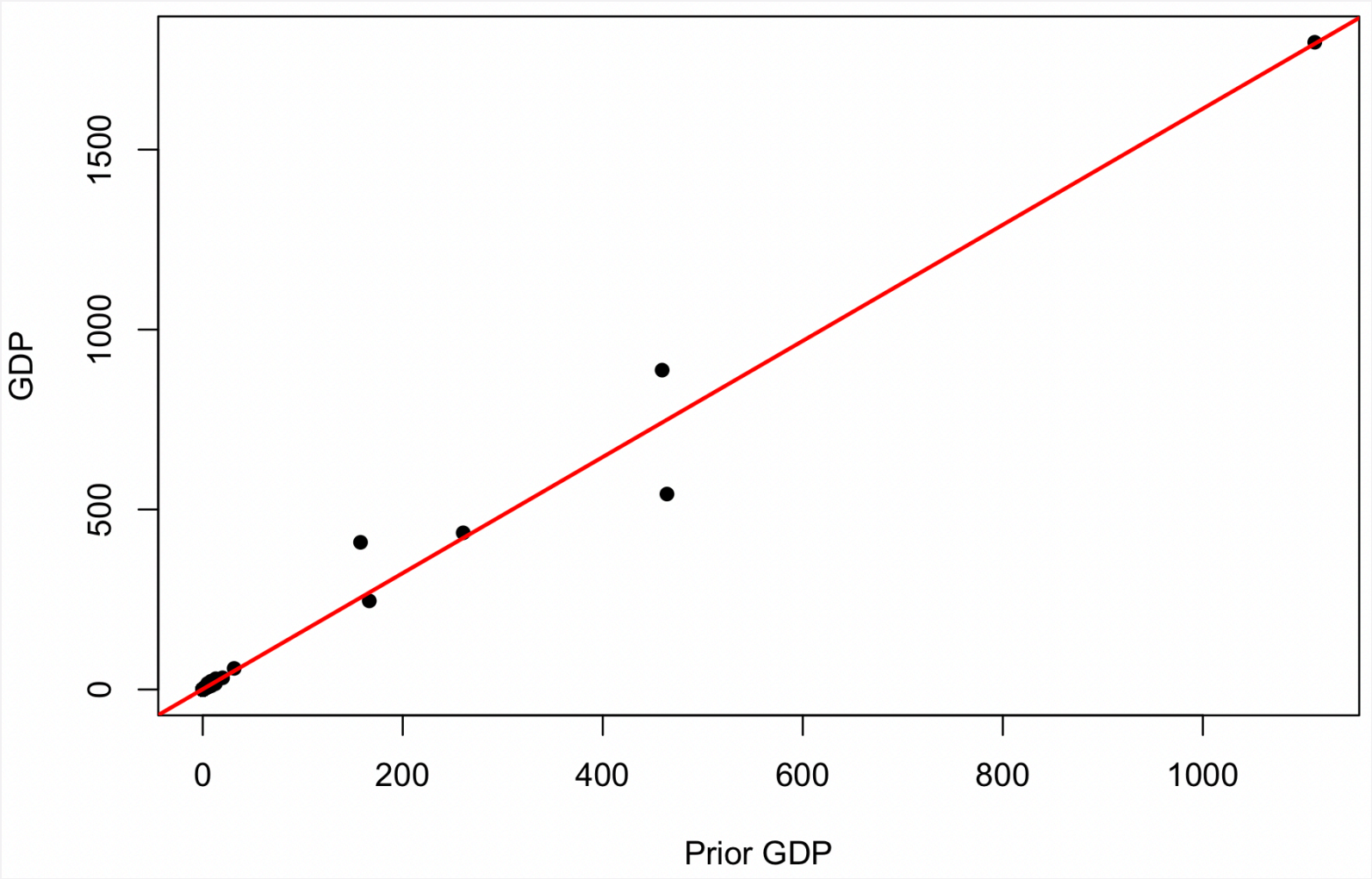
GDP



GDP

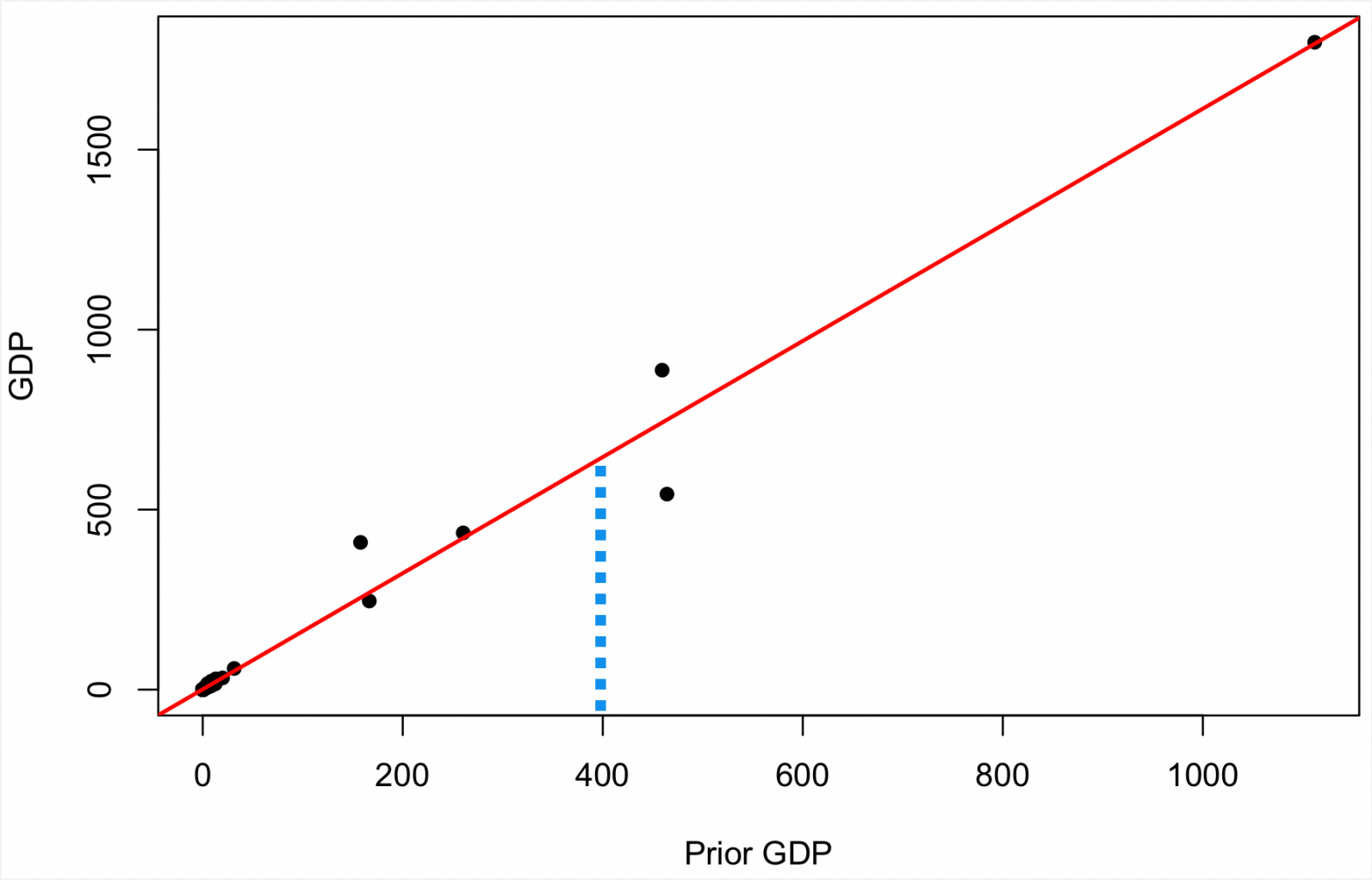


GDP



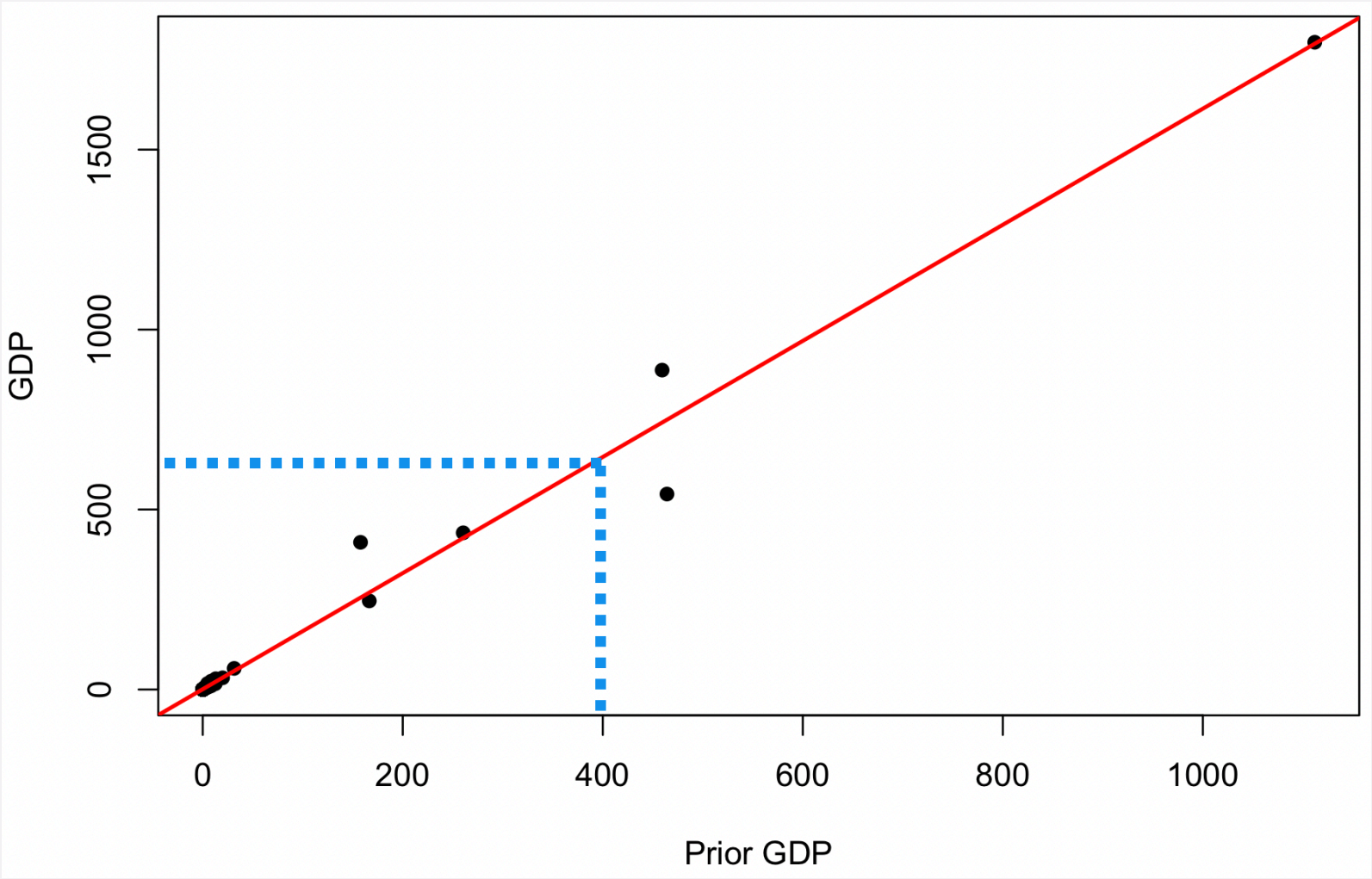
- **On average, how much higher is the current GDP of a country whose prior GDP was 400 instead of 600?**

GDP



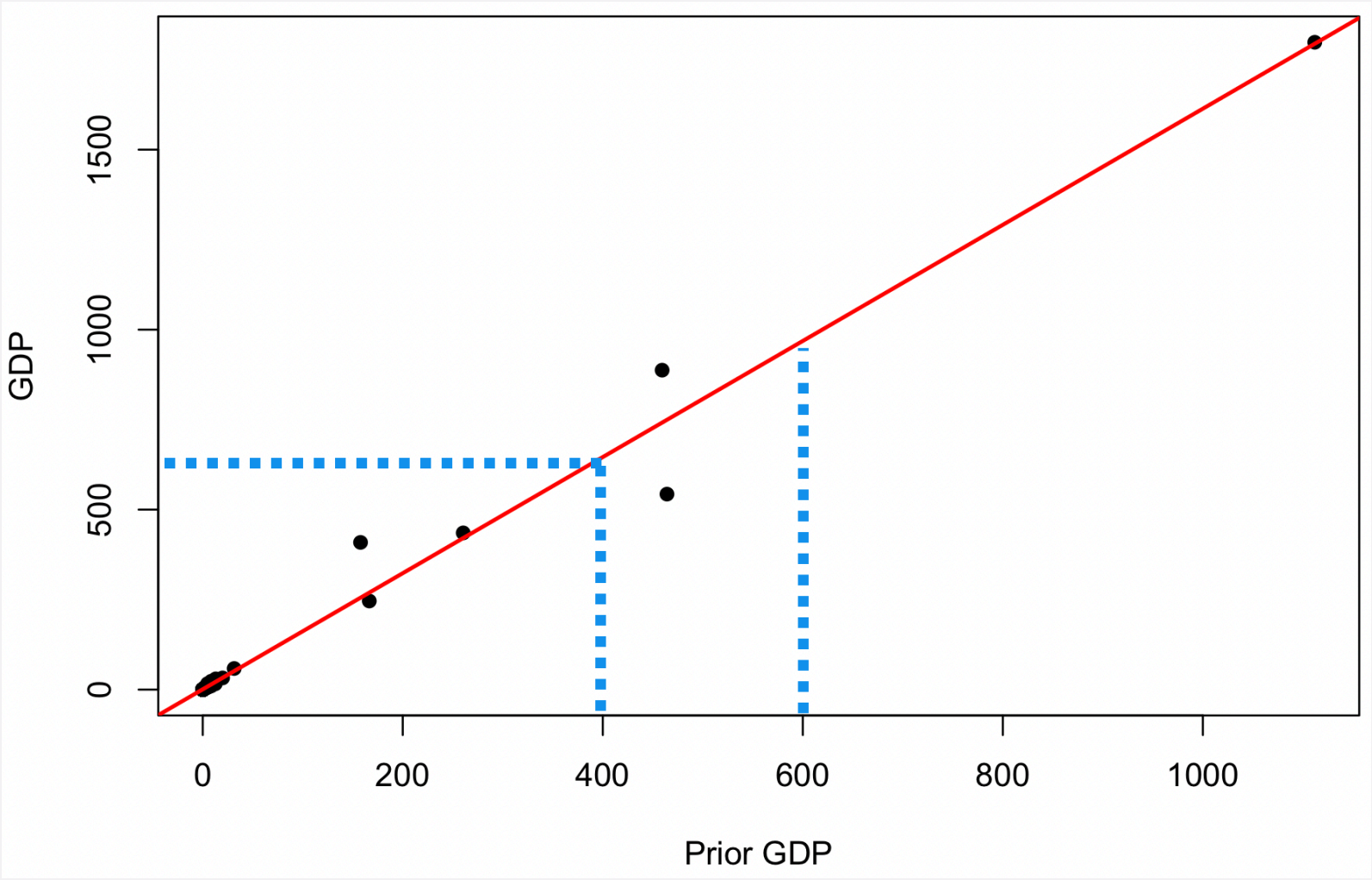
- **On average, how much higher is the current GDP of a country whose prior GDP was 400 instead of 600?**

GDP



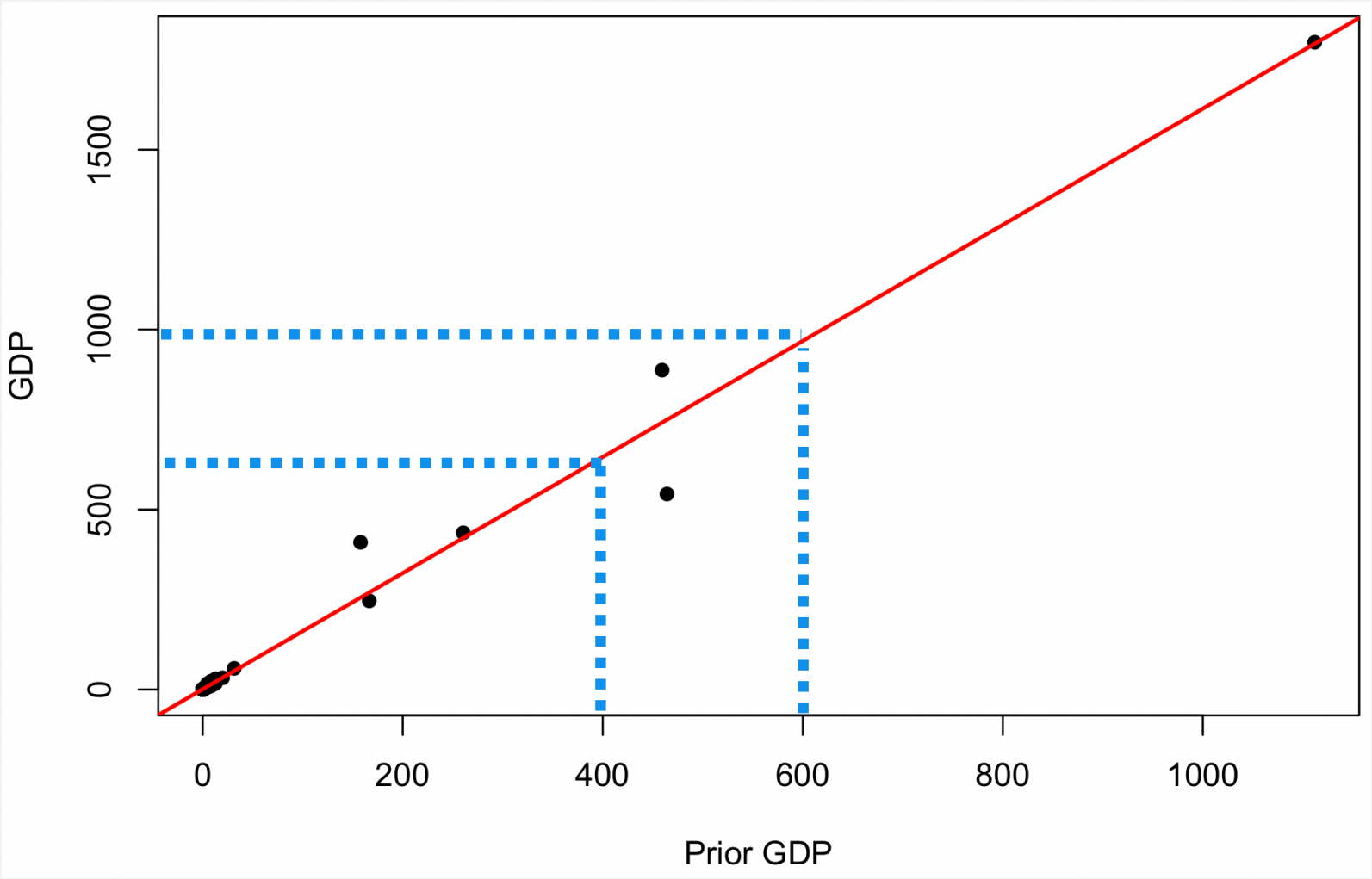
- **On average, how much higher is the current GDP of a country whose prior GDP was 400 instead of 600?**

GDP



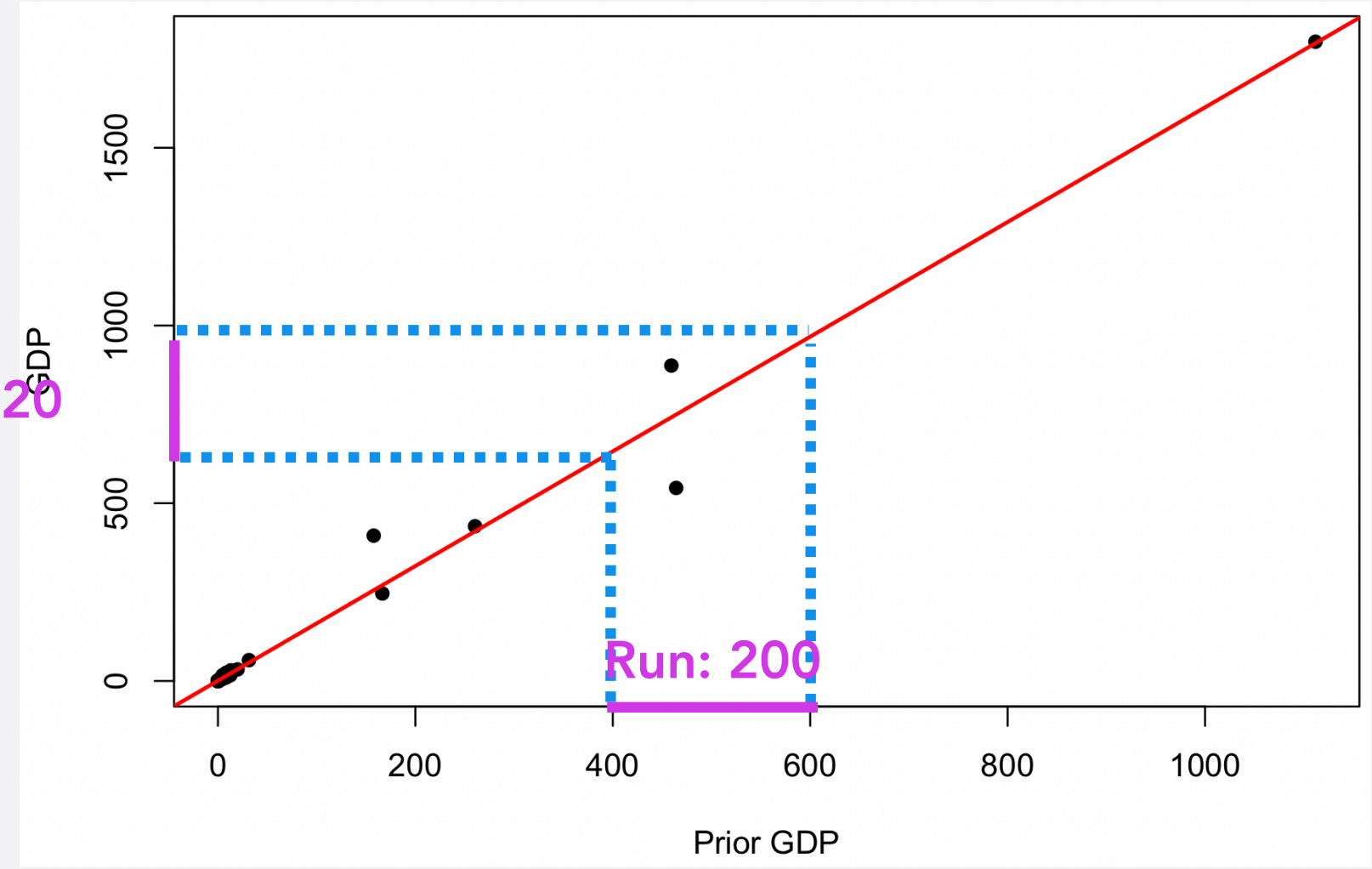
- **On average, how much higher is the current GDP of a country whose prior GDP was 400 instead of 600?**

GDP



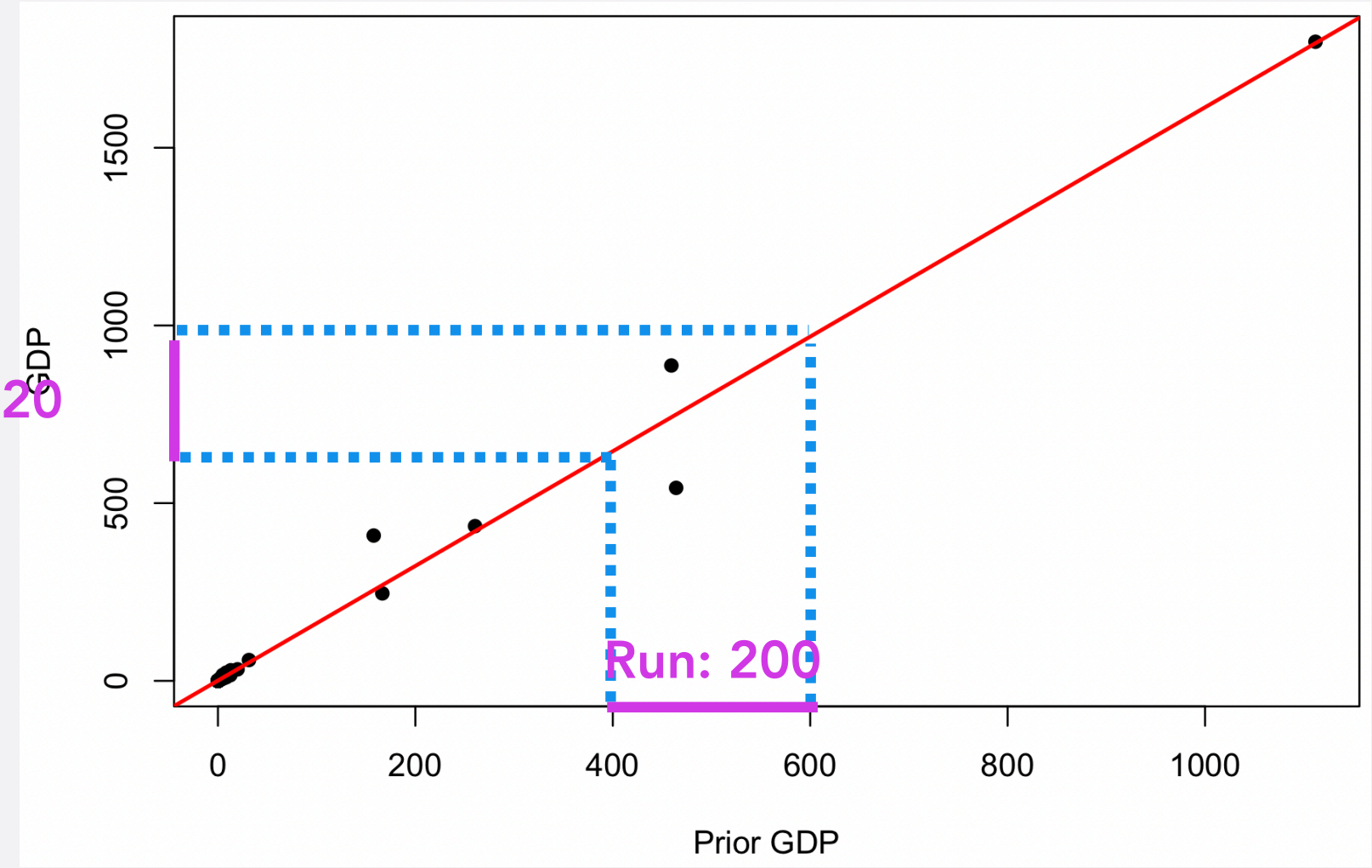
- **On average, how much higher is the current GDP of a country whose prior GDP was 400 instead of 600?**

GDP



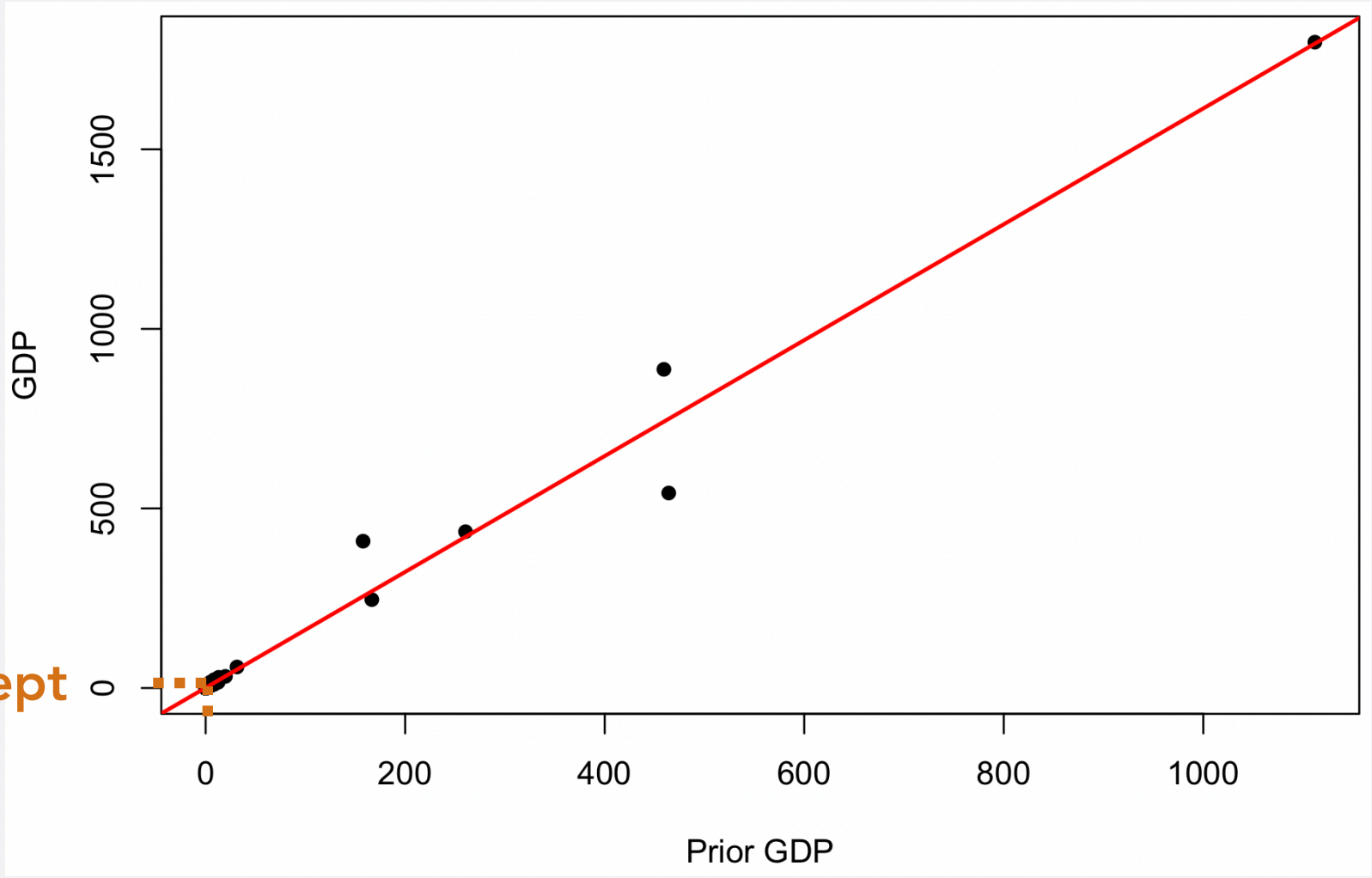
Slope=Rise over run=320/200=1.6

GDP



- For every one unit increase in prior GDP, current GDP is expected to increase by 1.6

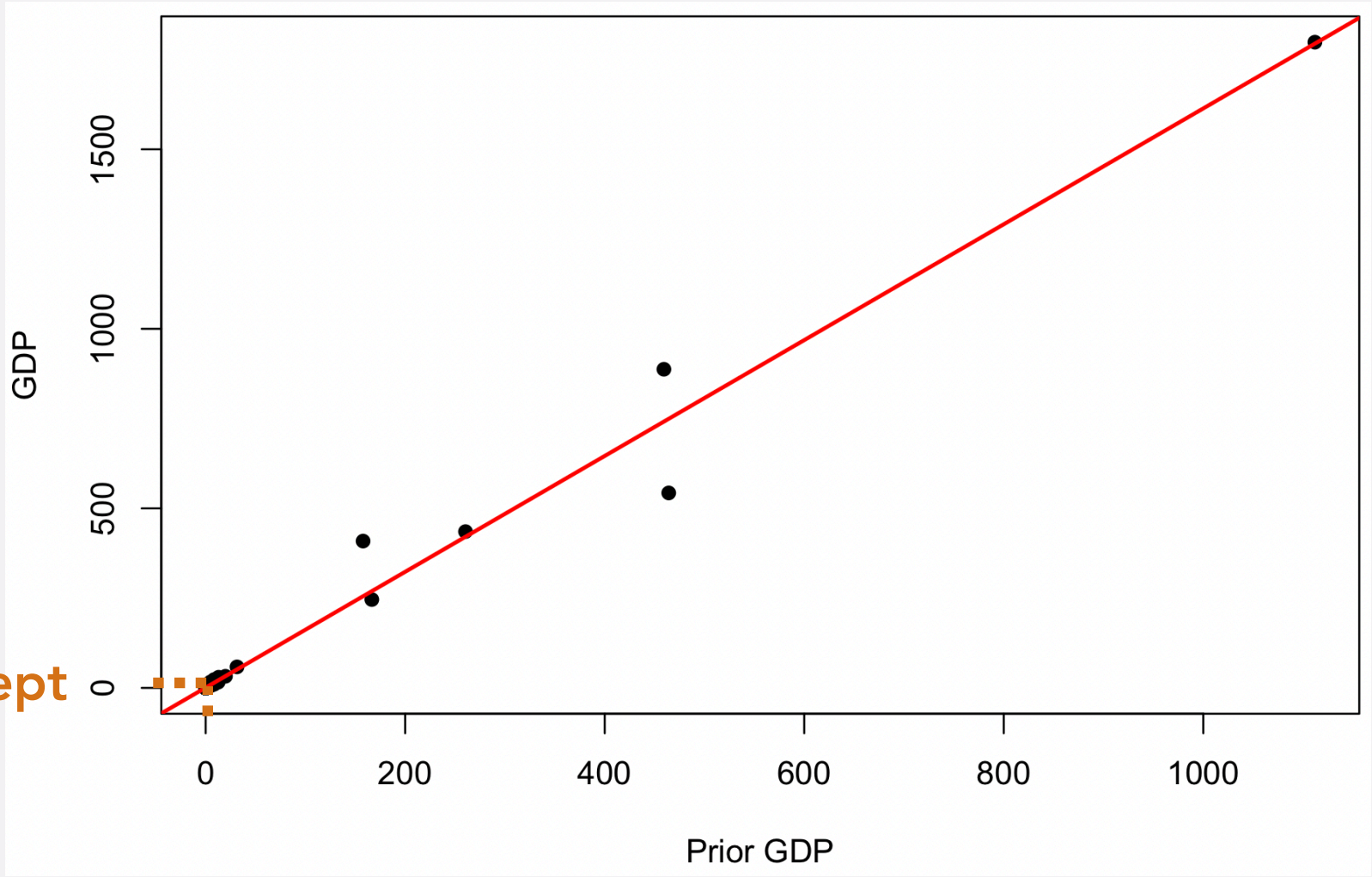
GDP



Intercept

Intercept=0.7

GDP



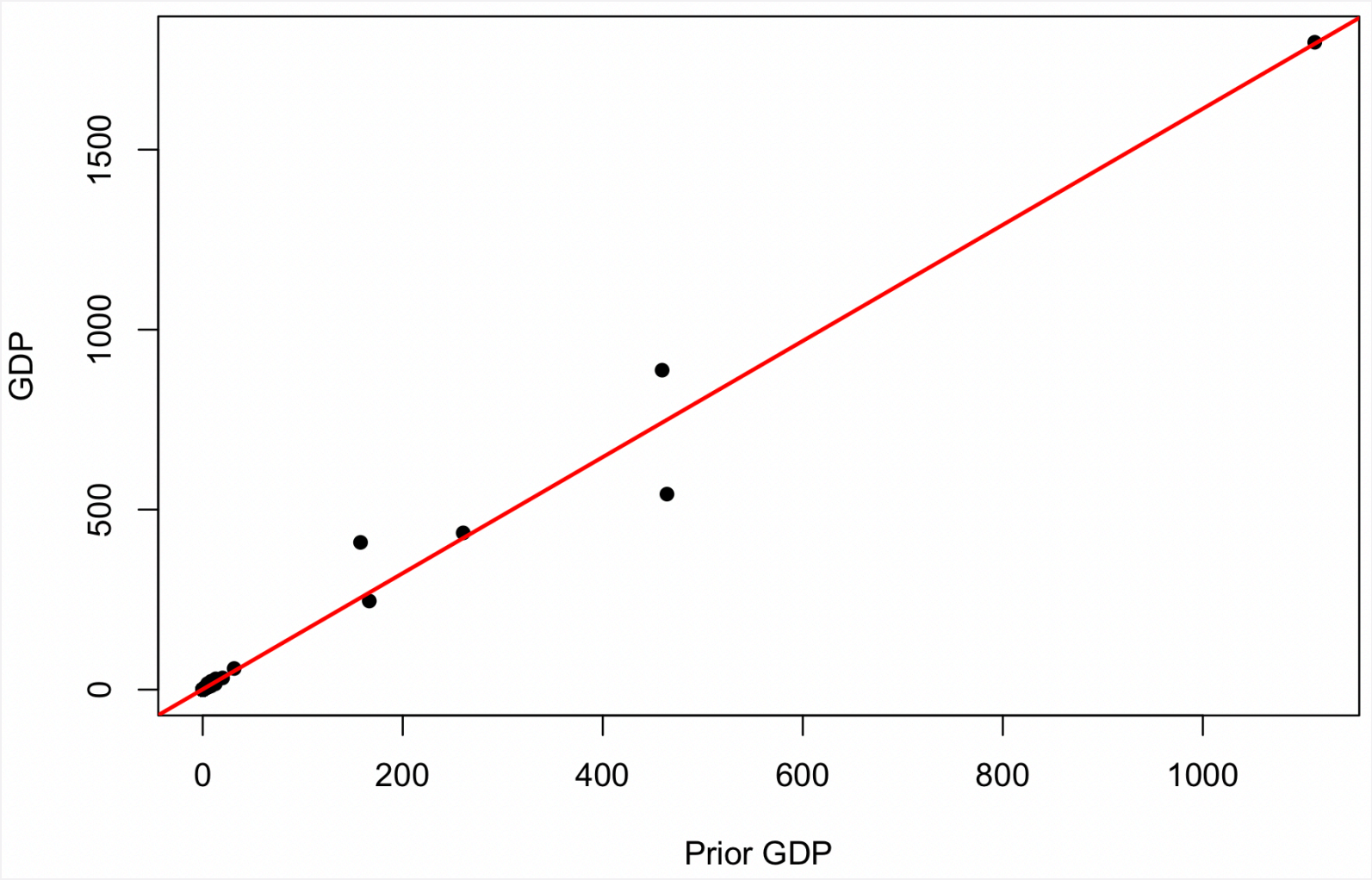
Intercept

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

LINEAR REGRESSION

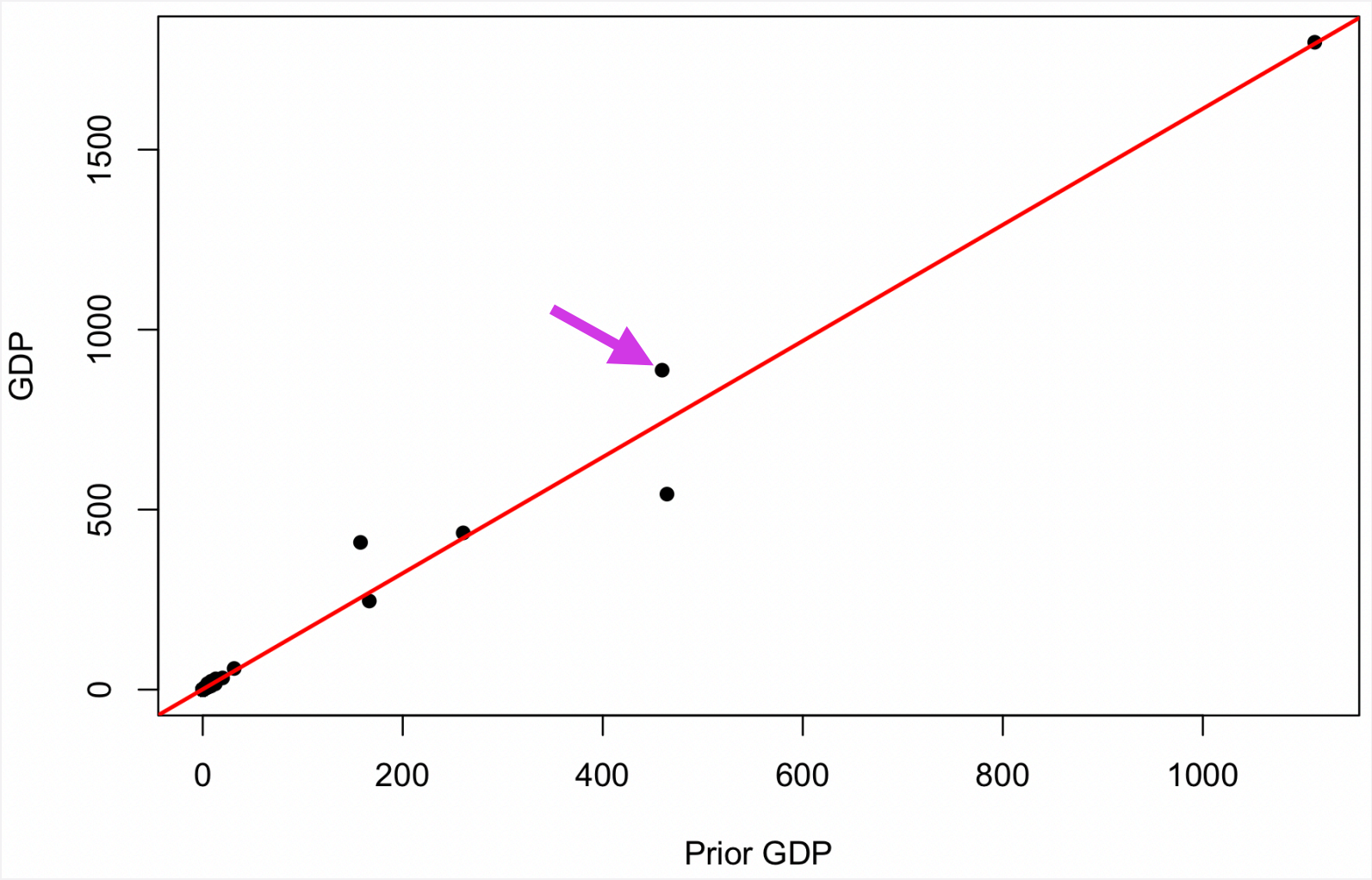
- Linear regression: Equation that tells us *direction* and *size* of relationship between independent variable (IV) and dependent variable (DV)
- DV = Intercept + Slope * IV + error

GDP



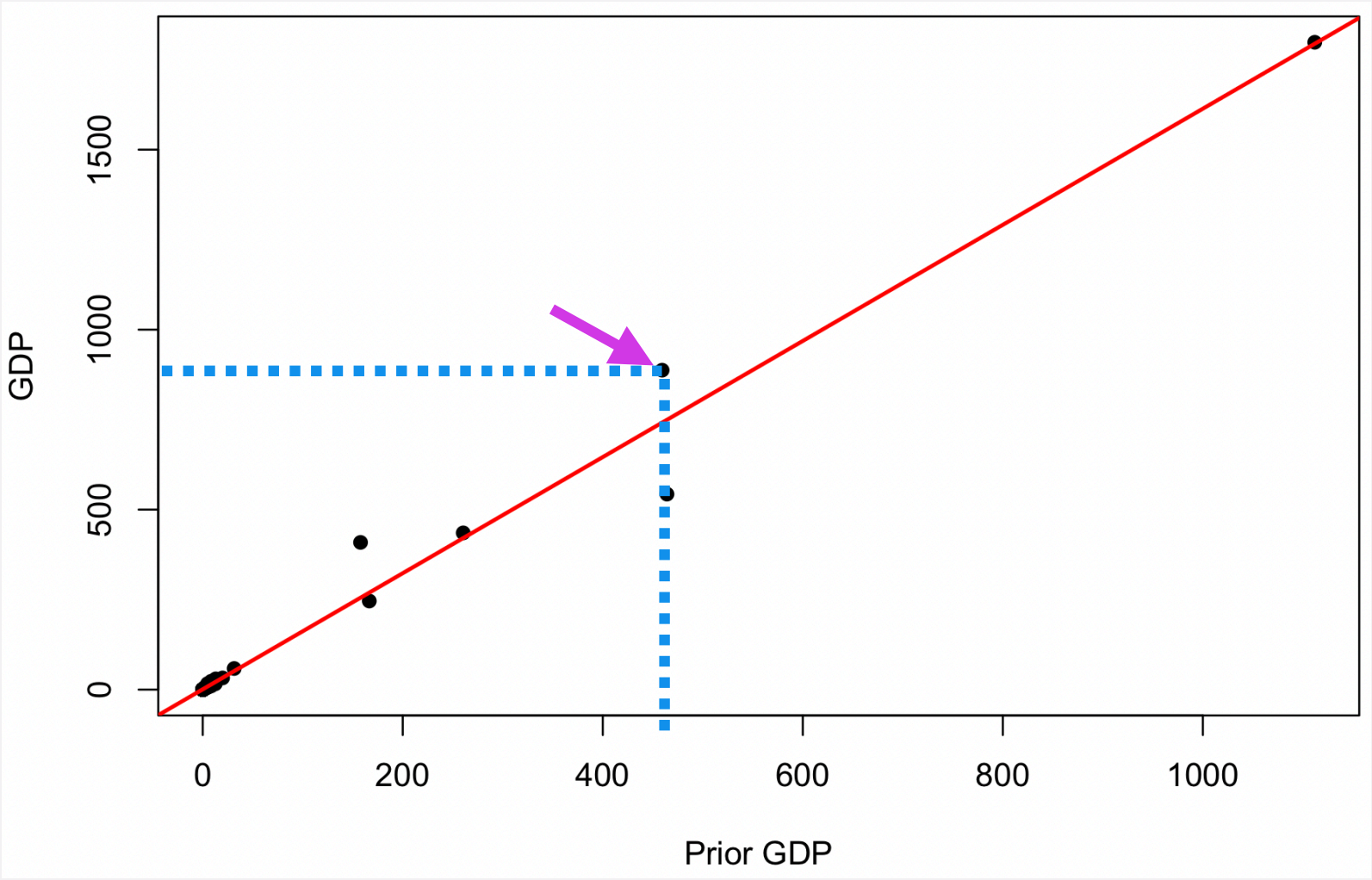
- GDP = 0.7 + 1.6 * Prior GDP + error

GDP



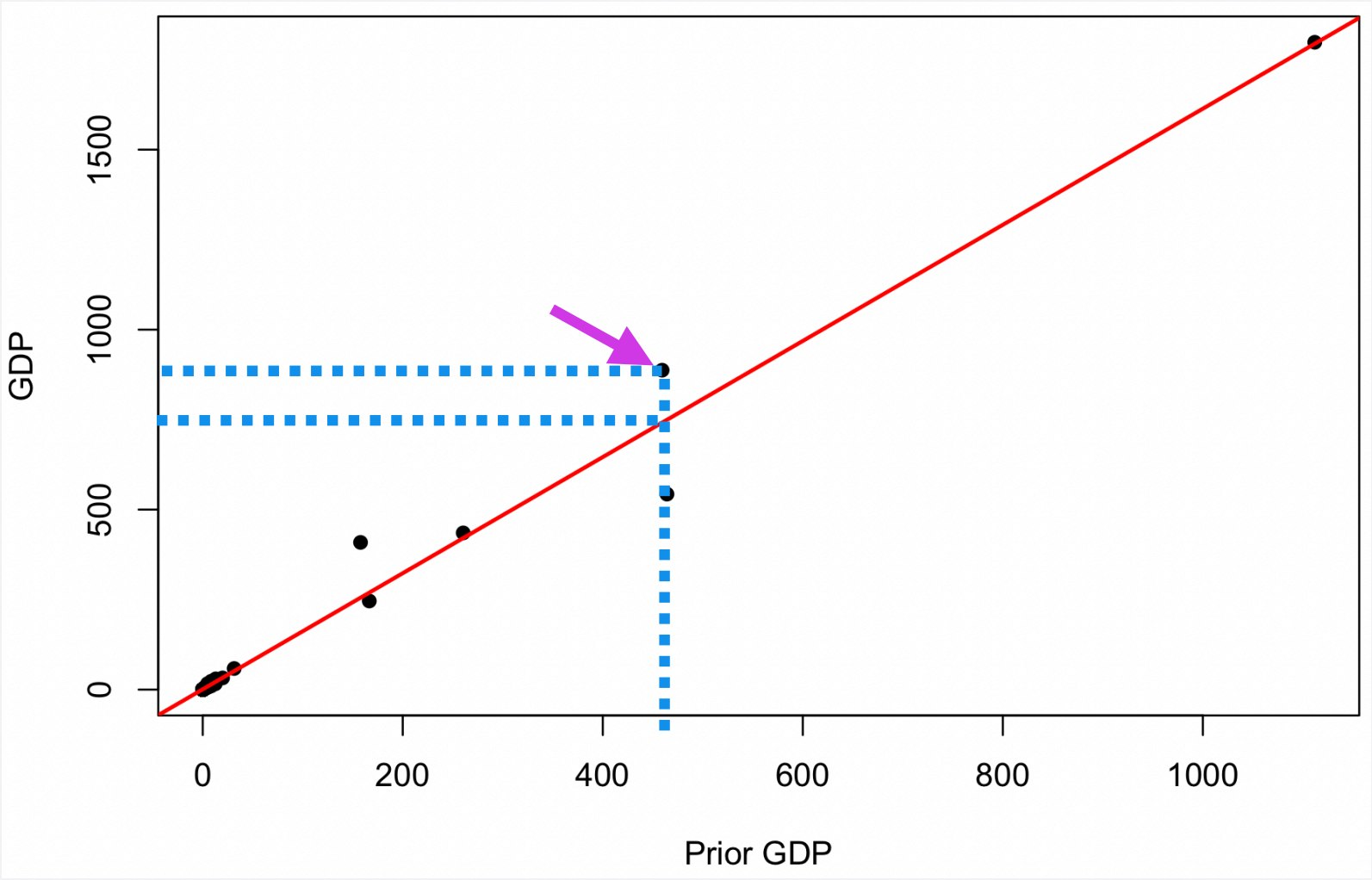
- GDP = 0.7 + 1.6 * Prior GDP + error

GDP



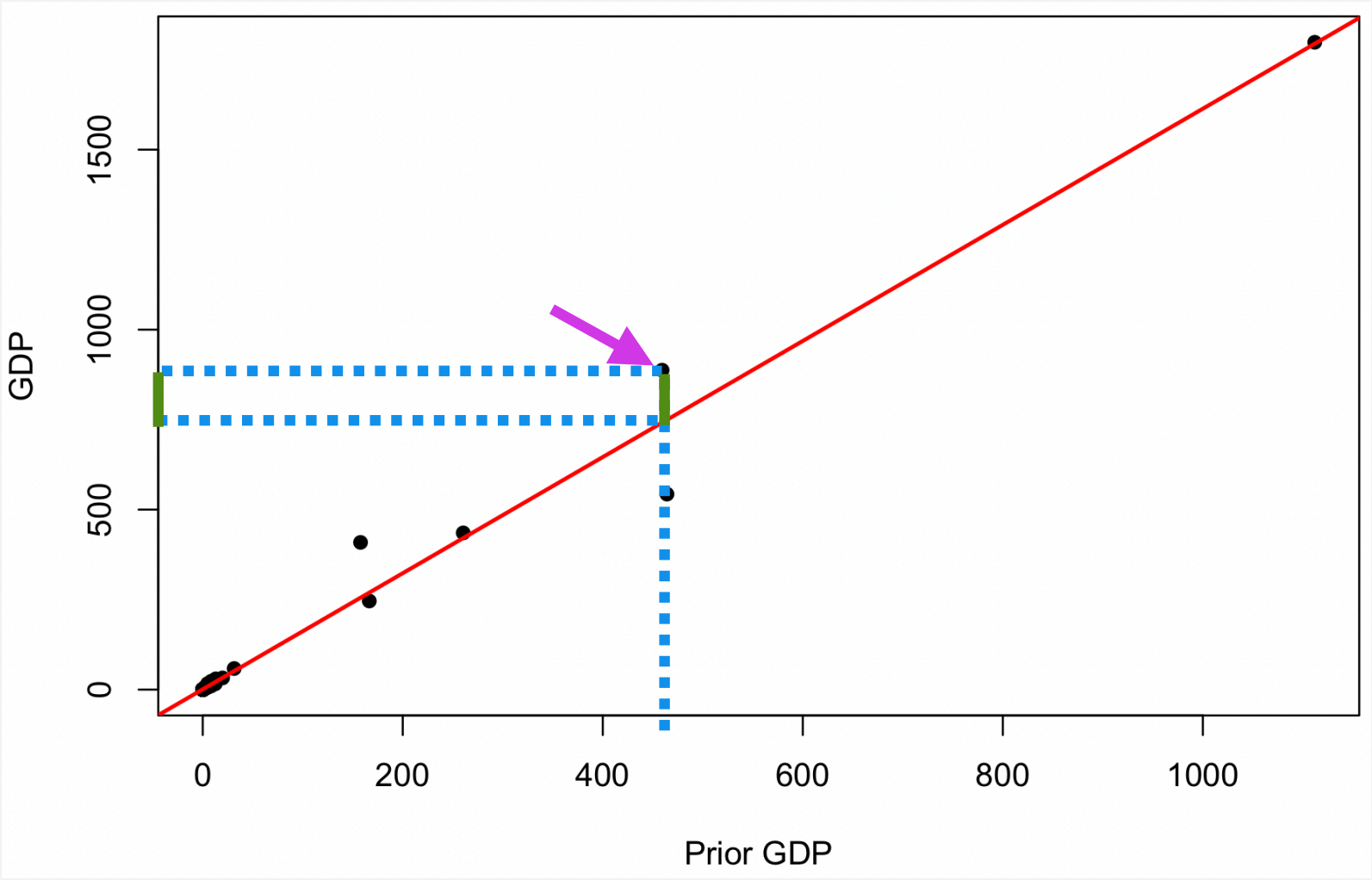
- GDP = 0.7 + 1.6 * Prior GDP + error

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

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

- pres08.csv