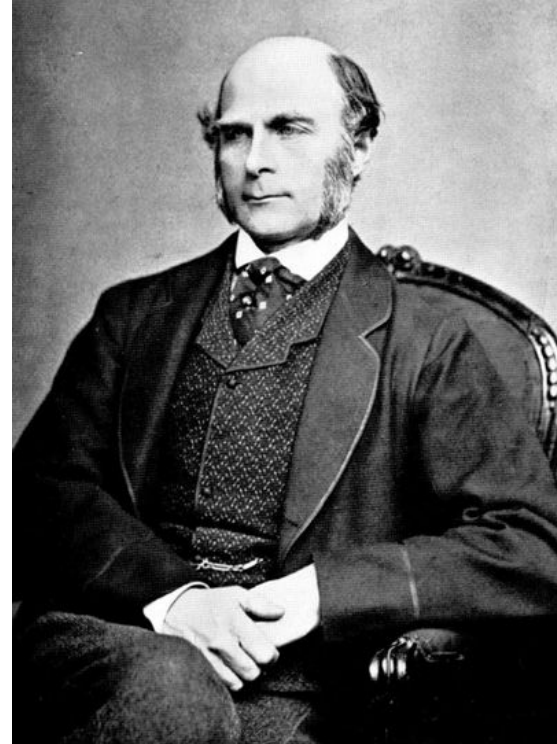


Introduction to Linear Regression

History

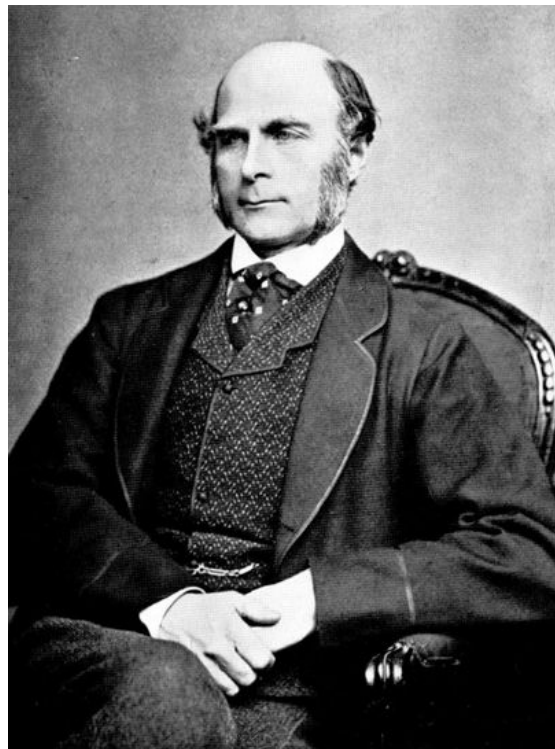
This all started in the 1800s with a guy named **Francis Galton**. Galton was studying the relationship between parents and their children. In particular, he investigated the relationship between the heights of fathers and their sons.



History

What he discovered was that a man's son tended to be roughly as tall as his father.

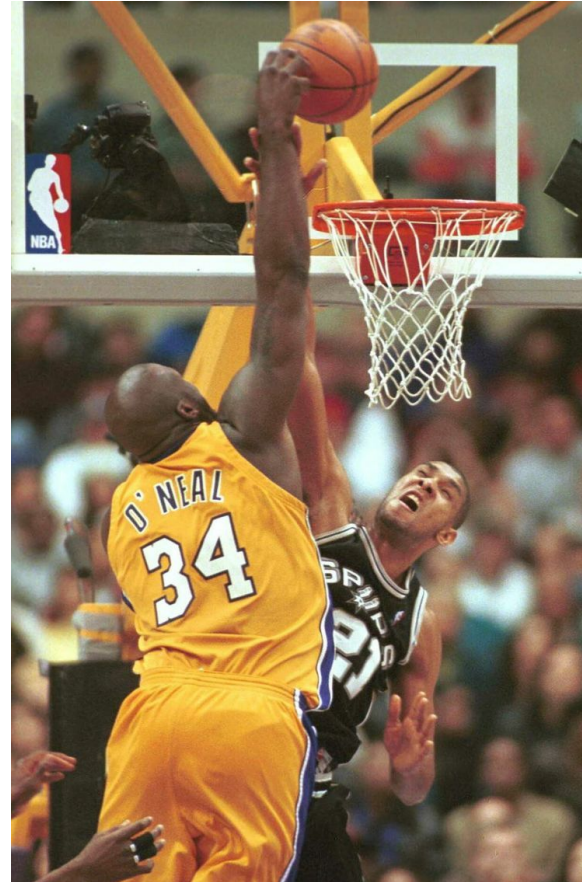
However Galton's breakthrough was that the son's height **tended to be closer to the overall average** height of all people.



Example

Let's take **Shaquille O'Neal** as an example. Shaq is really tall: 7ft 1in (2.2 meters).

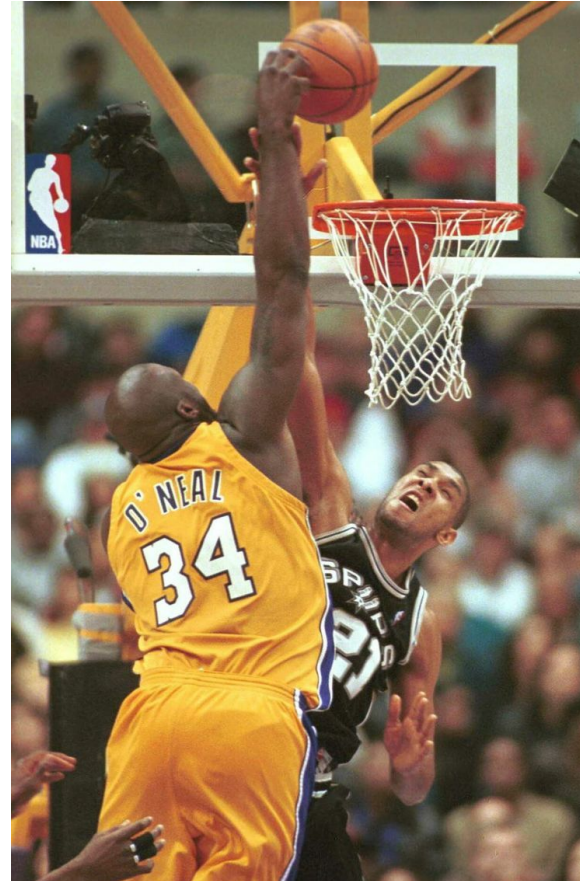
If Shaq has a son, chances are he'll be pretty tall too. However, Shaq is such an anomaly that there is also a very good chance that his son will be **not be as tall as Shaq**.



Example

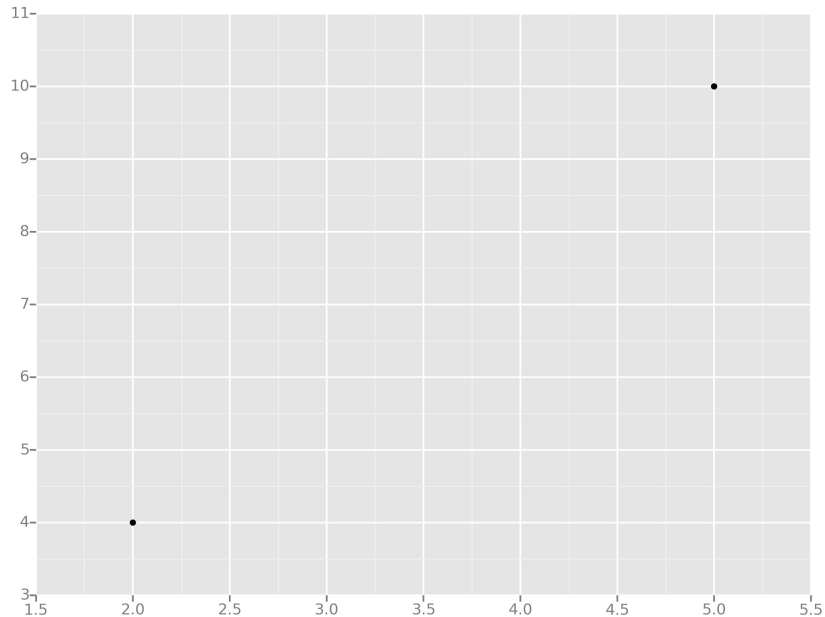
Turns out this is the case:
Shaq's son is pretty tall (6 ft 7 in), but not nearly as tall as his dad.

Galton called this phenomenon **regression**, as in "A father's son's height tends to regress (or drift towards) the mean (average) height."



Example

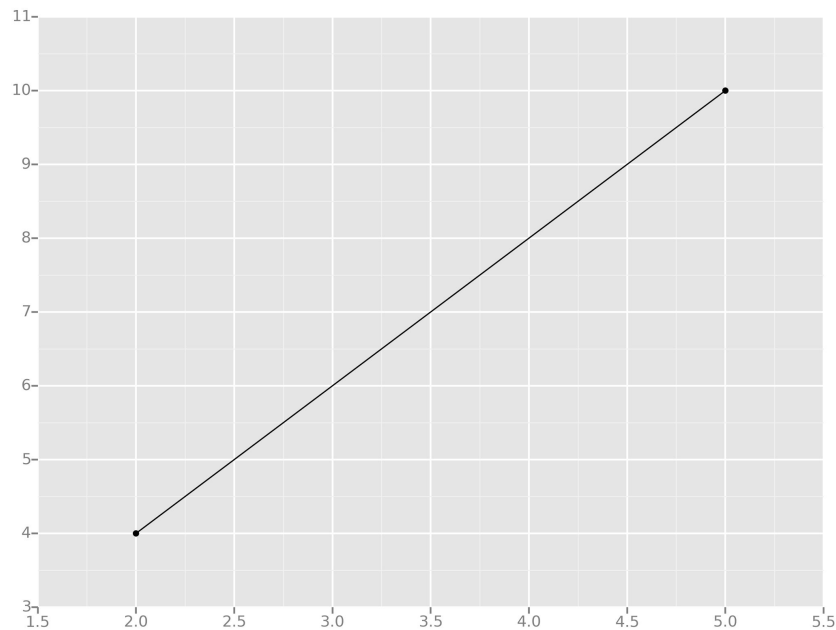
Let's take the simplest possible example:
calculating a regression
with only 2 data points.



Example

All we're trying to do when we calculate our regression line is draw a line that's as close to every dot as possible.

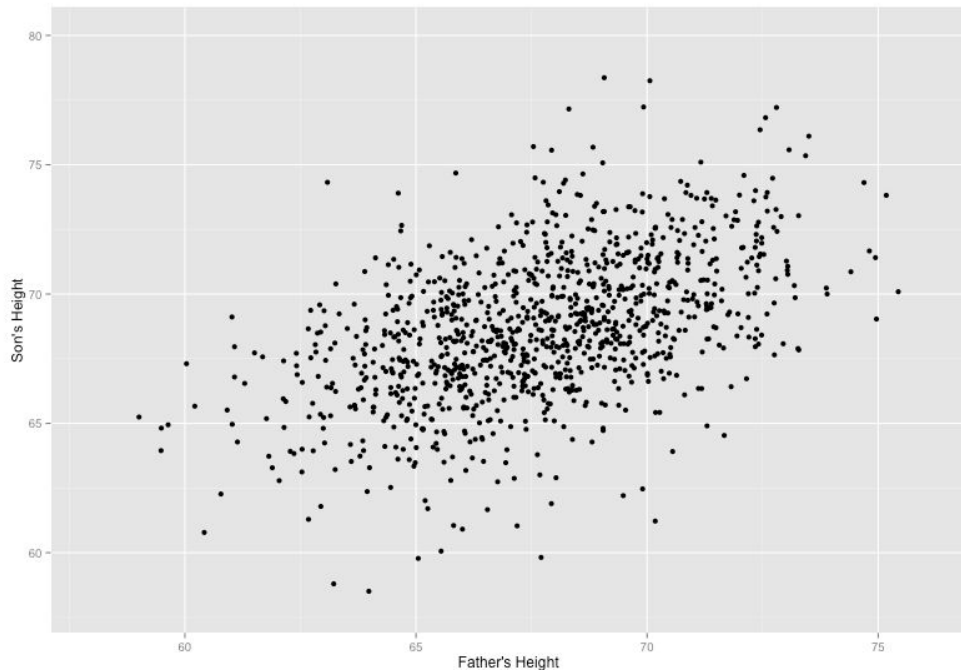
For classic linear regression, or "Least Squares Method", you only measure the closeness in the "up and down" direction



Example

Now wouldn't it be great if we could apply this same concept to a graph with more than just two data points?

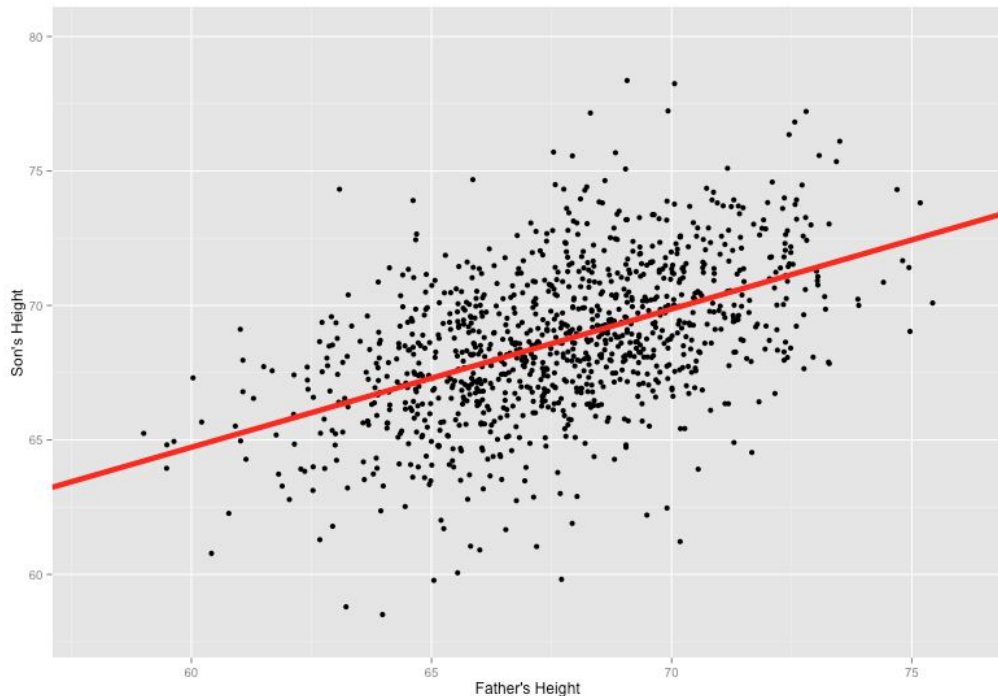
By doing this, we could take multiple men and their son's heights and do things like tell a man how tall we expect his son to be...before he even has a son!



Example

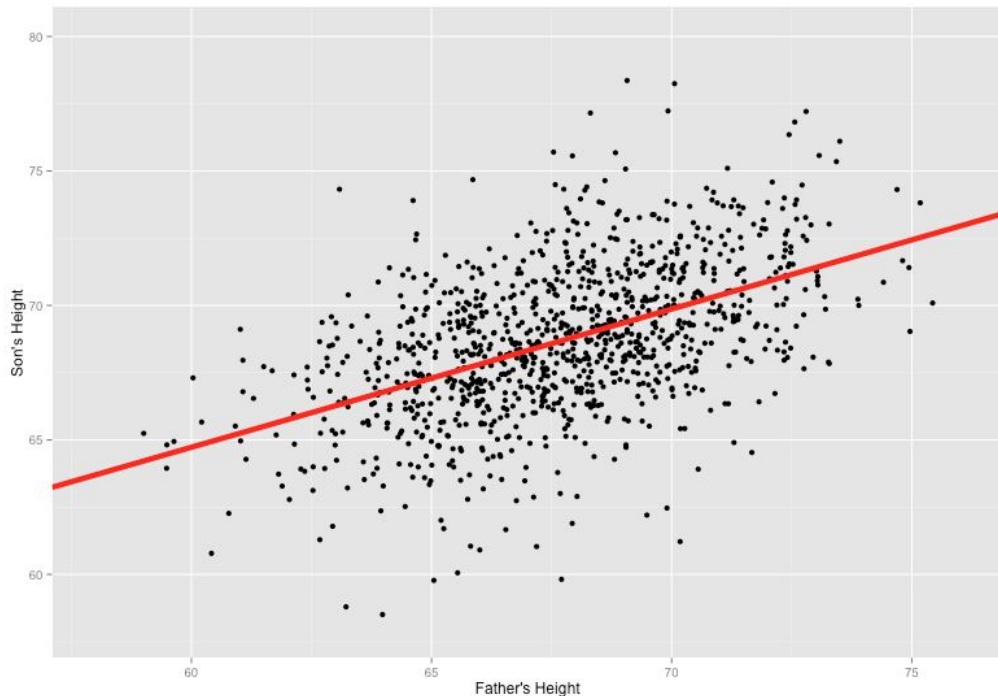
Our goal with linear regression is to **minimize the vertical distance** between all the data points and our line.

So in determining the **best line**, we are attempting to minimize the distance between **all** the points and their distance to our line.



Example

There are lots of different ways to minimize this, (sum of squared errors, sum of absolute errors, etc), but all these methods have a general goal of minimizing this distance.





COURSE NOTES: REGRESSION ANALYSIS

What is linear regression?

Regression analysis is one of the most widely used methods for prediction. Linear regression is probably the most fundamental machine learning method out there and a starting point for the advanced analytical learning path of every aspiring data scientist.

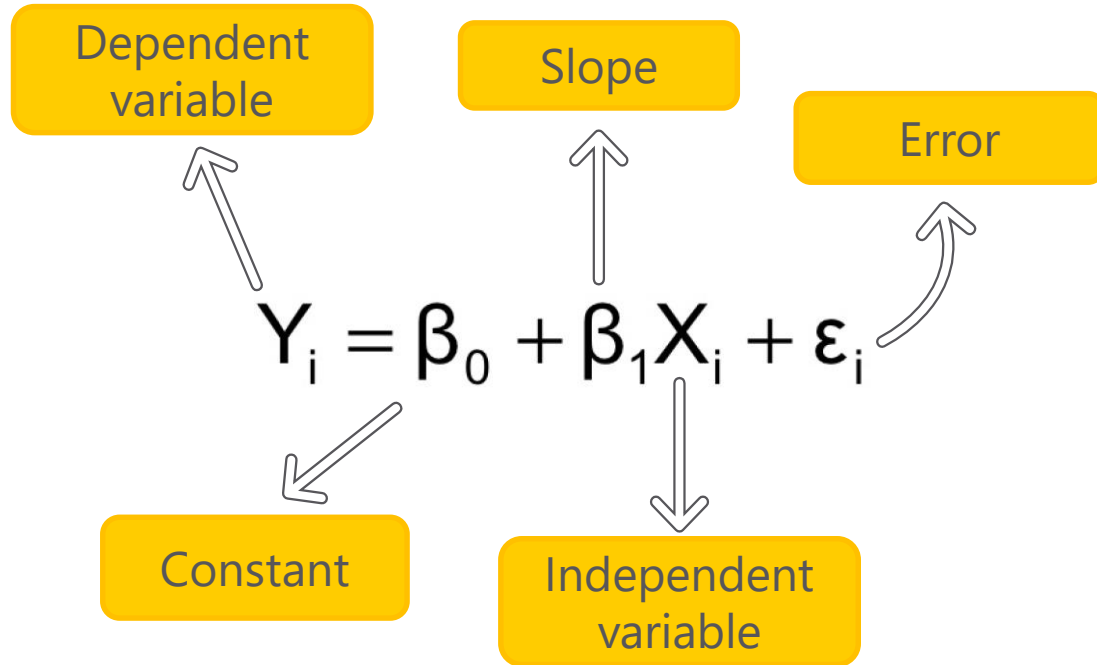
A linear regression is a linear approximation of a causal relationship between two or more variables.

Regression models are highly valuable, as they are one of the most common ways to make inferences and predictions. Apart from this, regression analysis is also employed to determine and assess factors that affect a certain outcome in a meaningful way.

As many other statistical techniques, regression models help us make predictions about the population based on sample data.

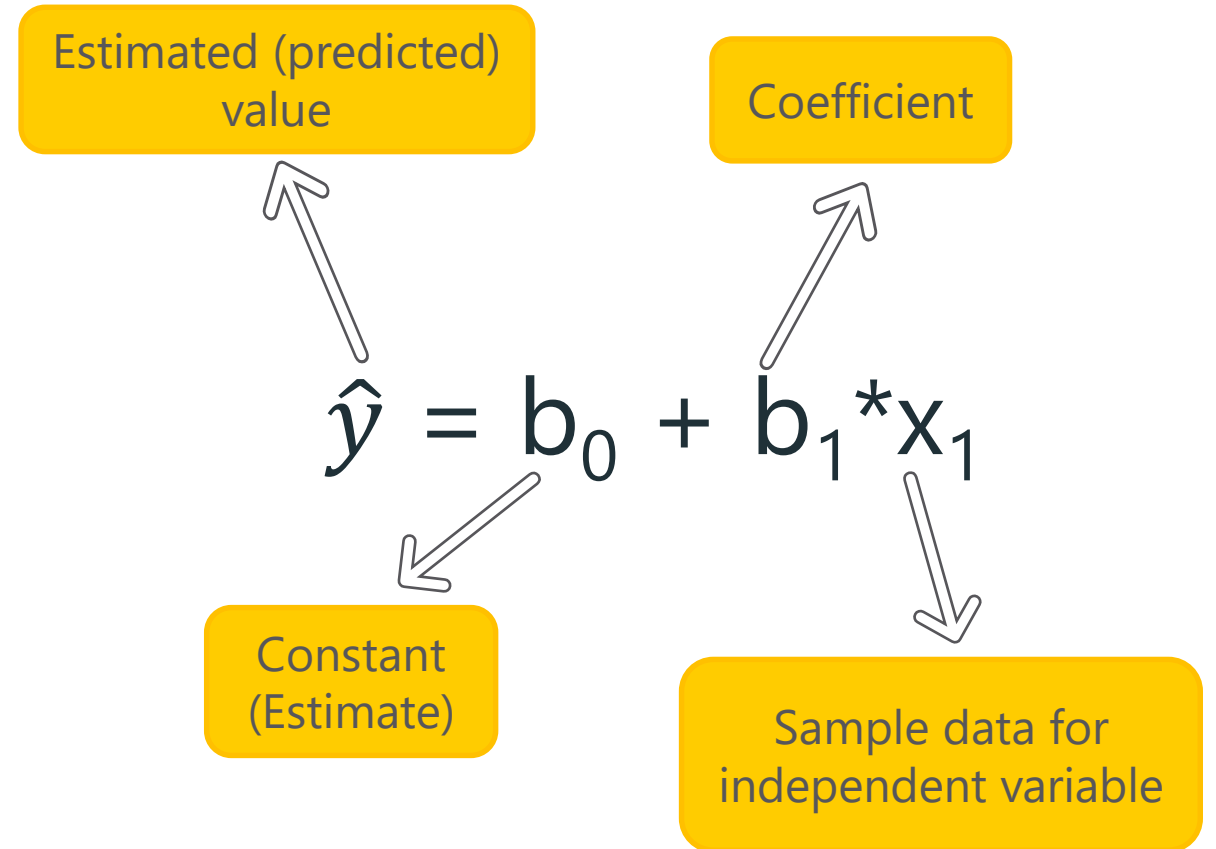


Linear regression model



Note: When we refer to the population models, we use **Greek letters**

Linear regression equation

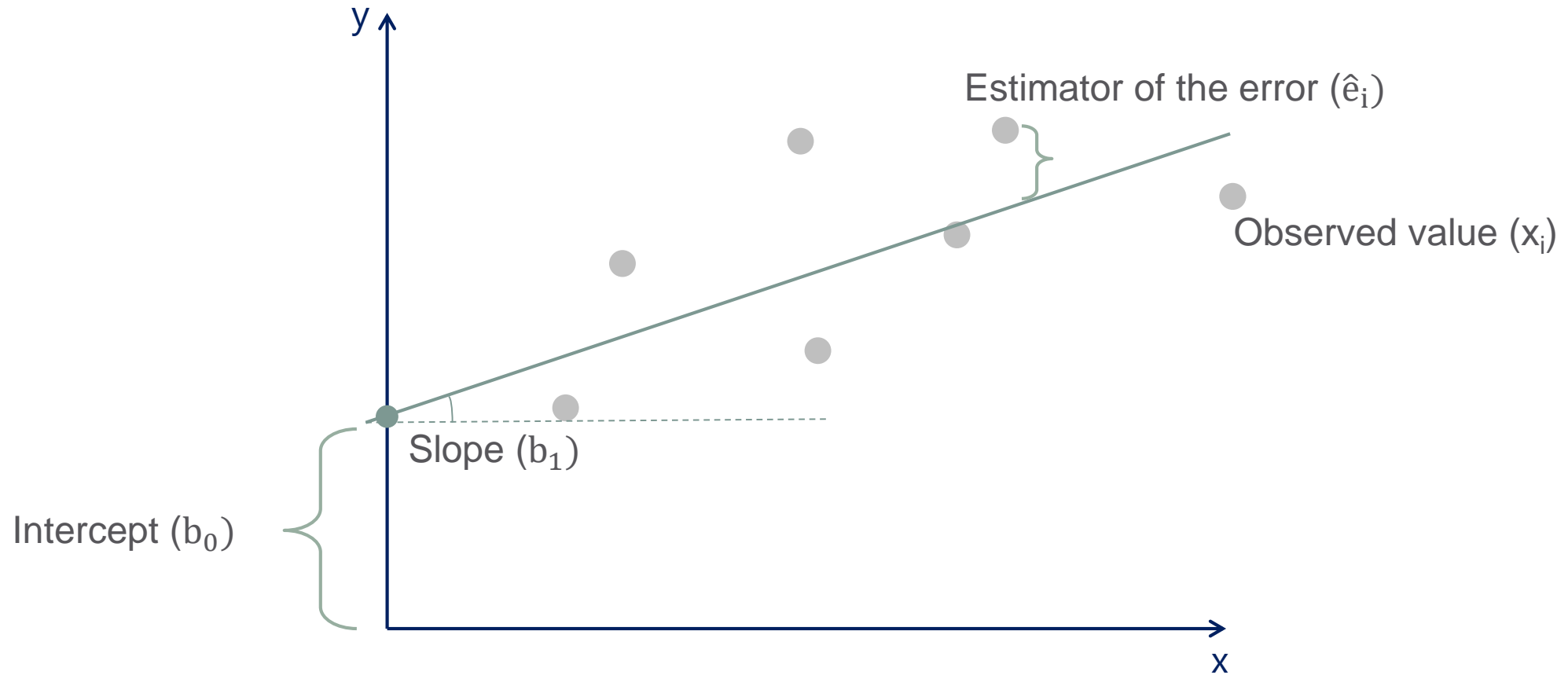


As many other statistical techniques, regression models help us make predictions about the population based on sample data.



Geometrical representation of linear regression

$$\hat{y}_i = b_0 + b_1 x_i$$



*On average the expected value of the error is 0, that is why it is not included in the regression equation

Correlation

vs

Regression

Represents the relationship
between two variables

Shows that two variables move
together (no matter in which direction)

Symmetrical w.r.t. the two variables:
 $\rho(x,y) = \rho(y,x)$

A single point (a number)

Represents the relationship
between two or more variables

Shows cause and effect (one variable
is affected by the other)

One way – there is always only one
variable that is causally dependent

A line (in 2D space)

Summary table and important regression metrics



Variability of the data, explained by the regression model
Range: [0;1]

Variability of the data, explained by the regression model, considering the number of independent variables
Range: <1; could be negative, but a negative number is interpreted as 0

The dependent variable, y ; This is the variable we are trying to predict

Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Fri, 22 Nov 2019	Prob (F-statistic):	7.20e-11
Time:	15:29:11	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

P-value for F-statistic; F-statistic evaluates the overall significance of the model (if at least 1 predictor is significant, F-statistic is also significant)

Coefficient of the intercept, b_0 ; sometimes we refer to this variable as constant or bias (as it 'corrects' the regression equation with a constant value)

	coef	std err	t	P> t	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

P-value of t-statistic; The t-statistic of a coefficient shows if the corresponding independent variable is significant or not

Coefficient of the independent variable i : b_i ; this is usually the most important metric – it shows us the relative/absolute contribution of each independent variable of our model

Omnibus:	12.839	Durbin-Watson:	0.950
Prob(Omnibus):	0.002	Jarque-Bera (JB):	16.155
Skew:	-0.722	Prob(JB):	0.000310
Kurtosis:	4.590	Cond. No.	3.29e+04

OLS assumptions

OLS (ordinary least squares) is one of the most common methods for estimating the linear regression equation. However, its simplicity implies that it cannot be always used. Therefore, all OLS regression assumptions should be met before we can rely on this method of estimation.

Linearity



$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$

The specified model must represent a linear relationship

No endogeneity



$$\sigma_{X\varepsilon} = 0 : \forall x, \varepsilon$$

The independent variables shouldn't be correlated with the error term.

Normality and homoscedasticity



$$\varepsilon \sim N(0, \sigma^2)$$

The variance of the errors should be consistent across observations.

No autocorrelation



$$\sigma_{\varepsilon_i \varepsilon_j} = 0 : \forall i \neq j$$

No identifiable relationship should exist between the values of the error term.

No multicollinearity



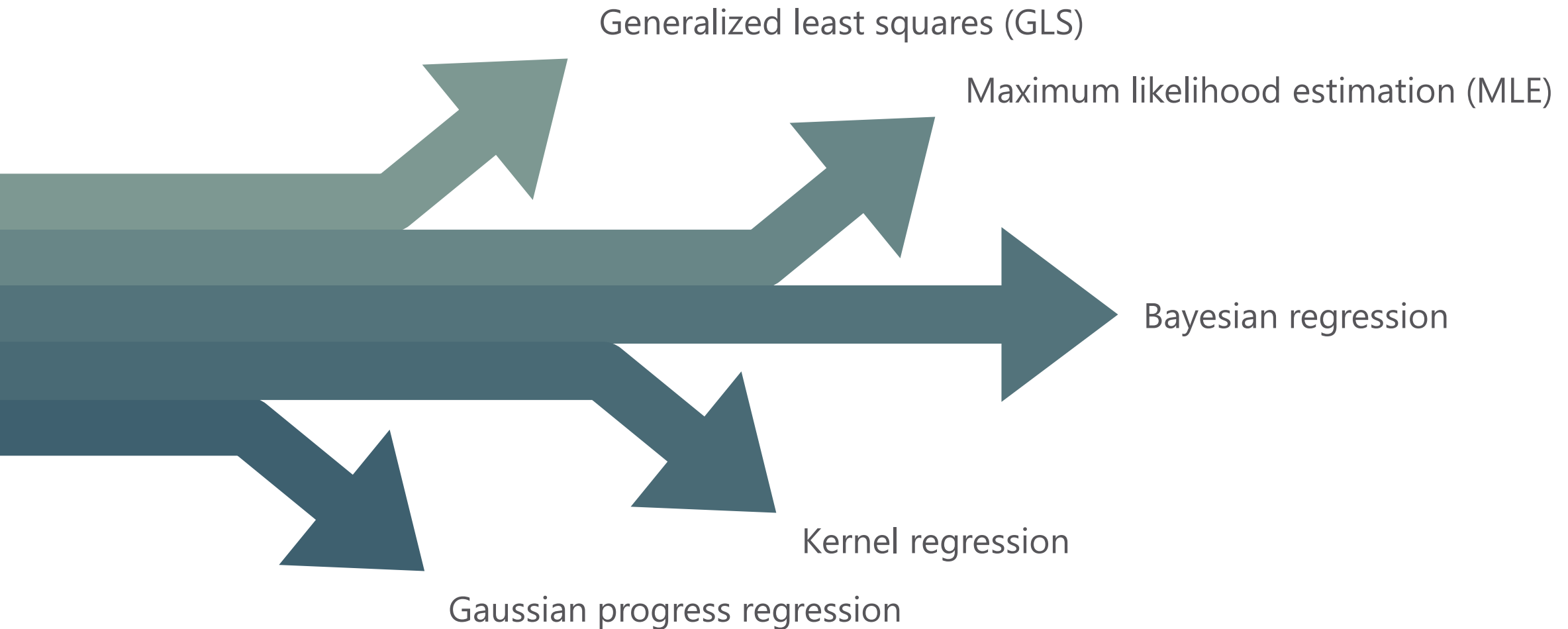
$$\rho_{x_i x_j} \approx 1 : \forall i, j; i \neq j$$

No predictor variable should be perfectly (or almost perfectly) explained by the other predictors.



Other methods for finding the regression line

OLS (ordinary least squares) is just the beginning. OLS is the simplest, although often sufficient method to estimate the regression line. In fact, there are more complex methods that are more appropriate for certain datasets and problems.



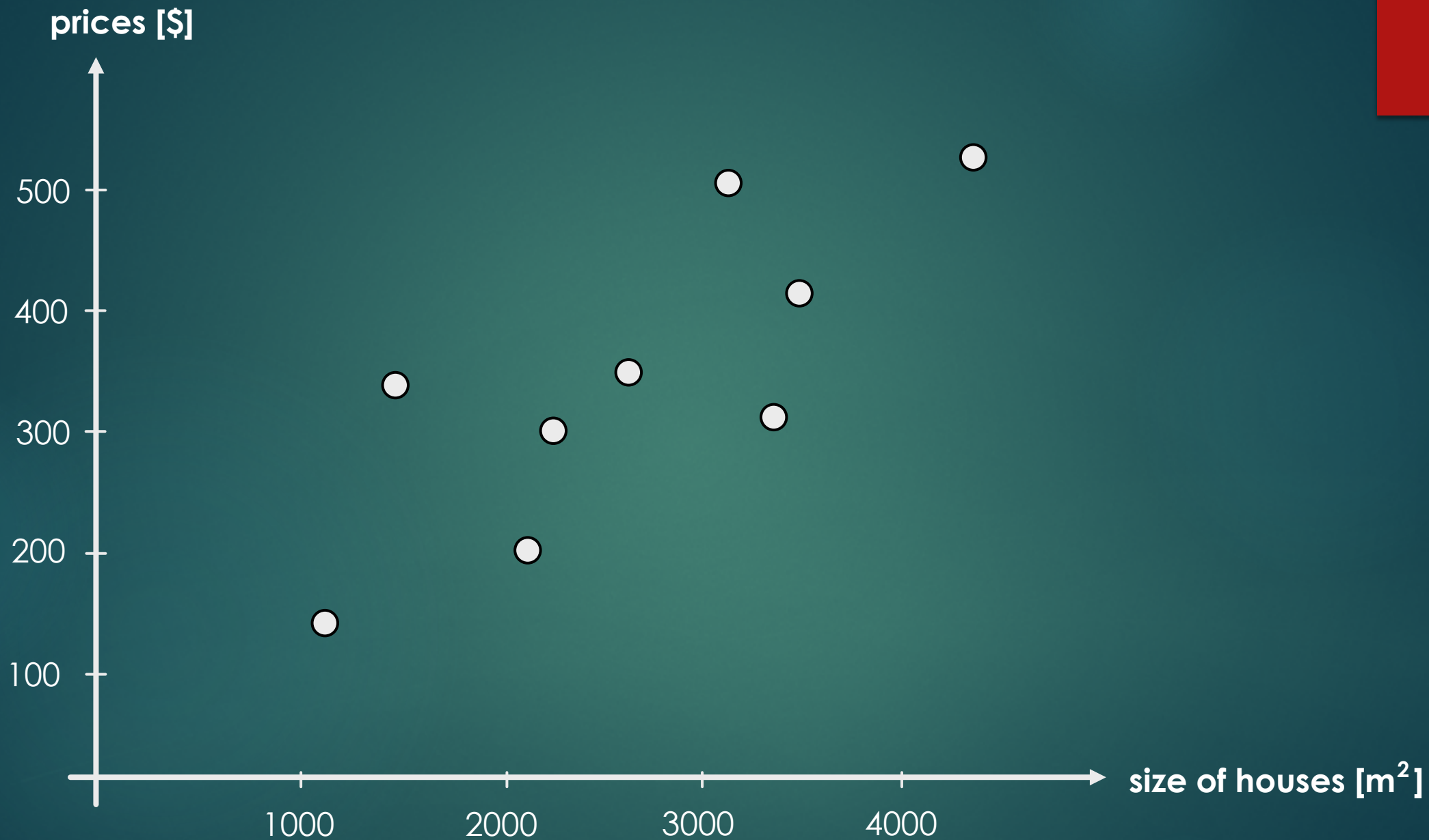


MACHINE LEARNING

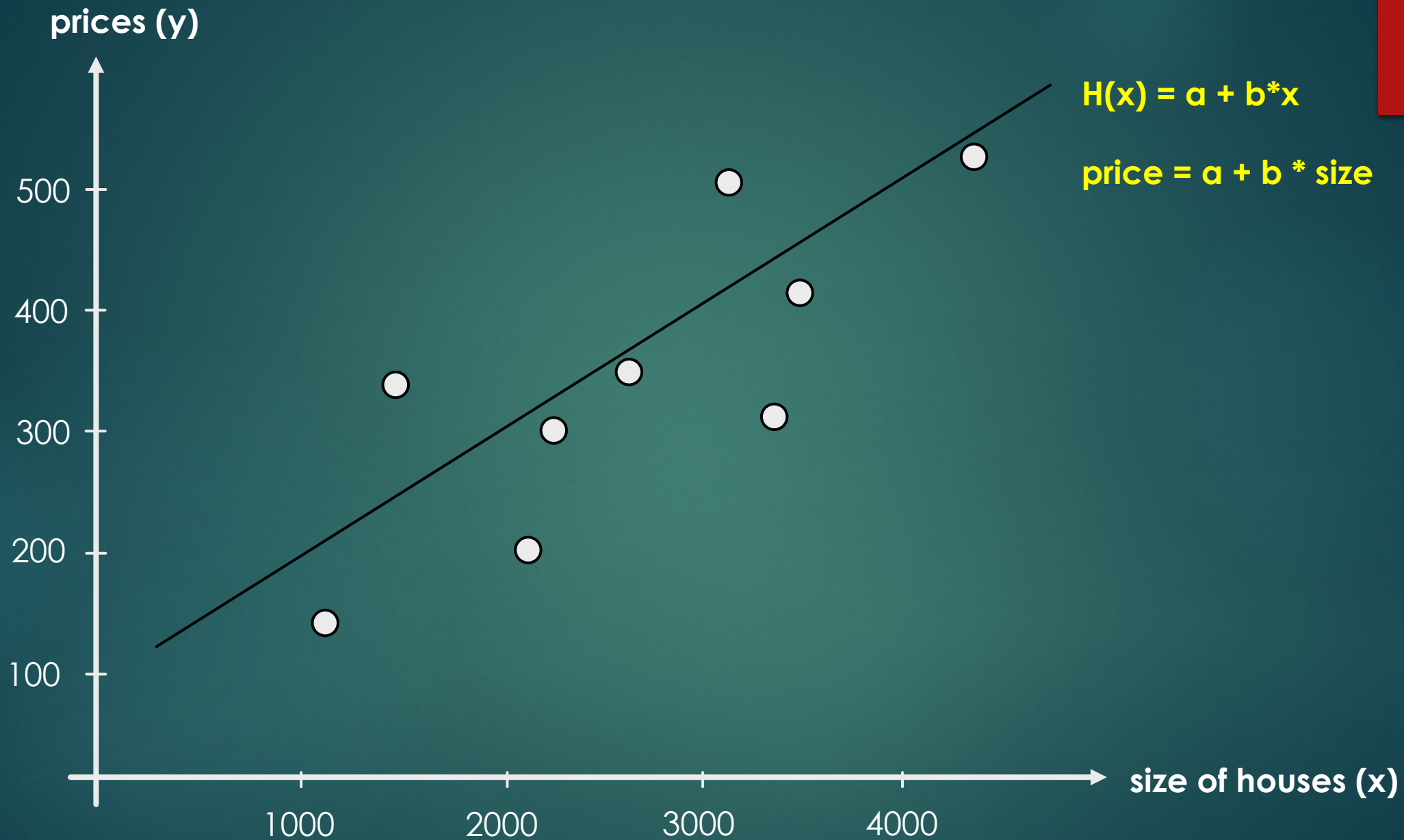
LINEAR REGRESSION

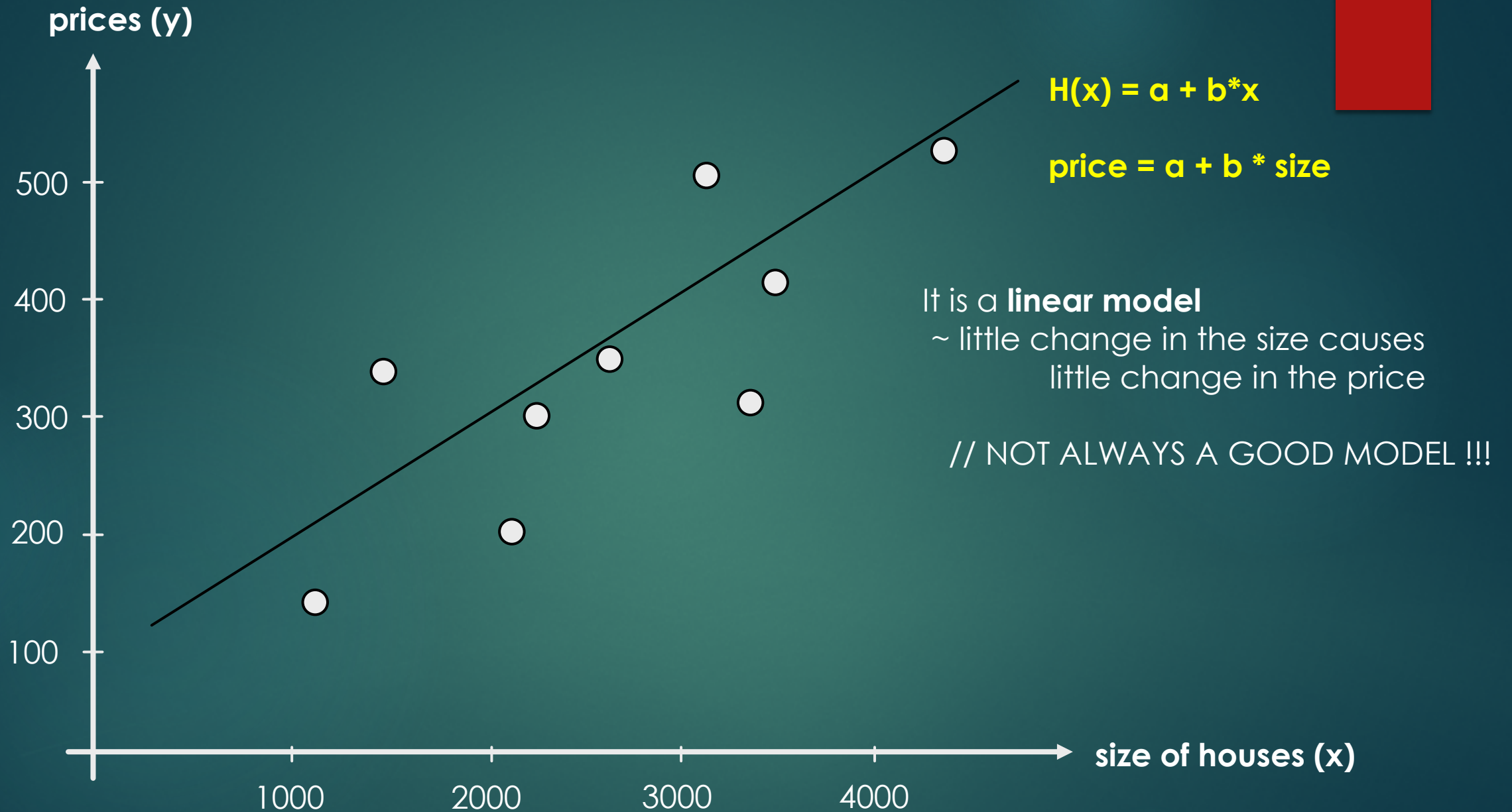
Linear regression

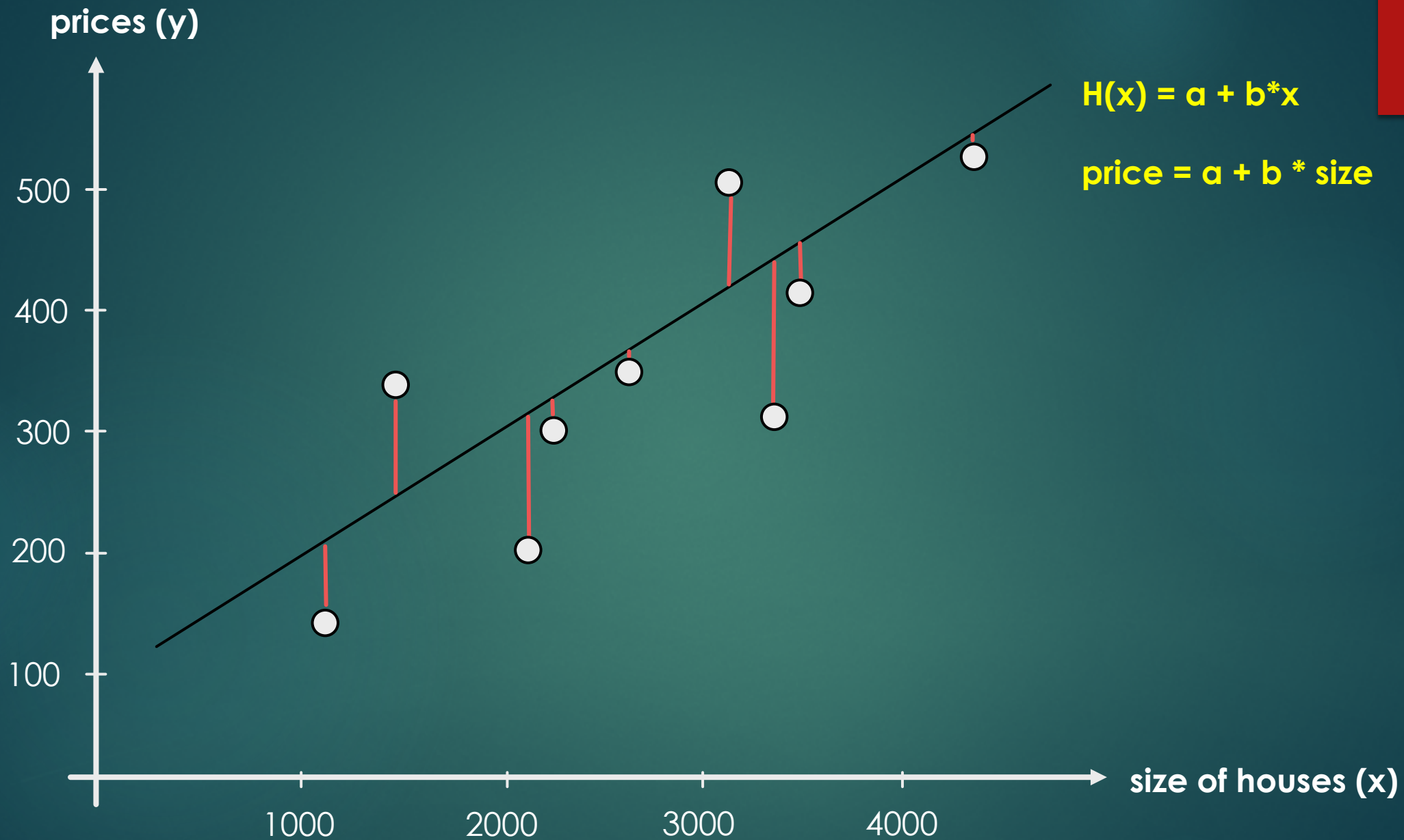
- ▶ It is an approach for modelling the relationship between scalar dependent variable y and one or more explanatory variables \underline{x}
- ▶ If we have one explanatory variable → **simple linear regression**
- ▶ For example: we want to approximate the price of houses if we know the size // single explanatory variable
- ▶ If we have several explanatory variable → **multiple linear regression**
- ▶ For example: we want to approximate the price of houses if we know the size, number of bedrooms ...
- ▶ We use linear predictor functions → this is why it is **LINEAR** regression



X [m*m]	Y [\$]
1120	112
1523	345
2102	198
2230	305
2600	372
3200	550
3409	302
3689	420
4460	578







Linear regression

- ▶ We should choose **a** and **b** parameters so that **H(x)** is close to **y** for our training examples **(x,y)** !!!
- ▶ So: minimize $[H(x) - y]^2$
- ▶ We have to choose the parameters **a** and **b** to make it a minimum
- ▶ We can solve it with **gradient descent**, in higher dimensions it is very efficient
- ▶ Or we can solve it with **design matrices** if there are not much features: matrix operations are expensive in higher dimensions

this is the prediction from
our linear model

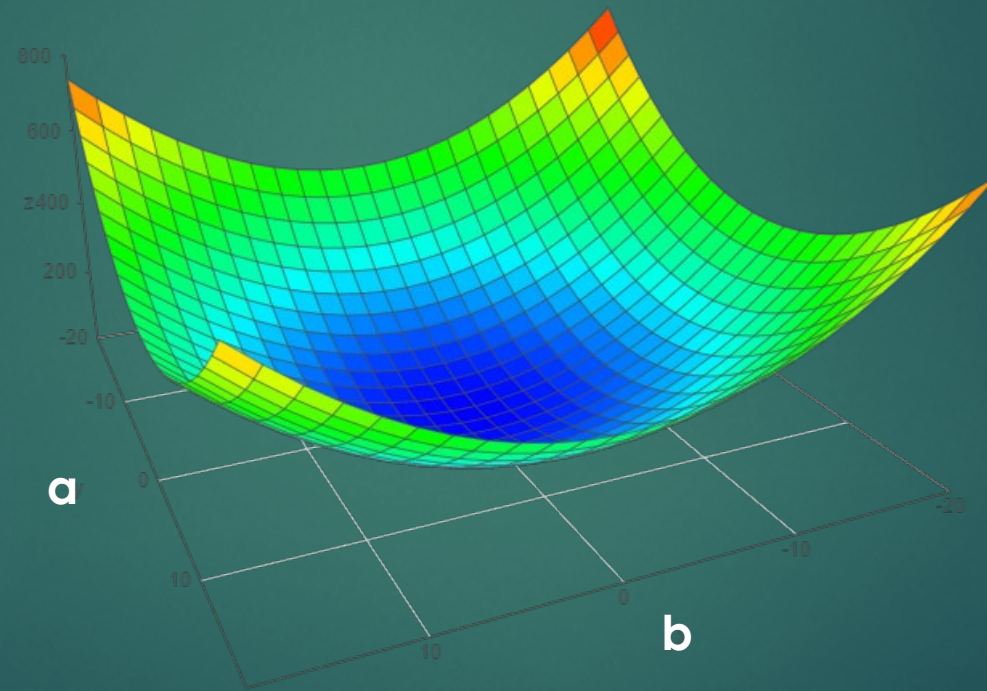
$$\min_{a,b} [H(x) - y]^2$$

we want to find the minimum by
tuning these parameters

this is the value we know from
the training data
// supervised learning !!!

Gradient descent

$\min C(a,b)$



Gradient descent

gradient descent

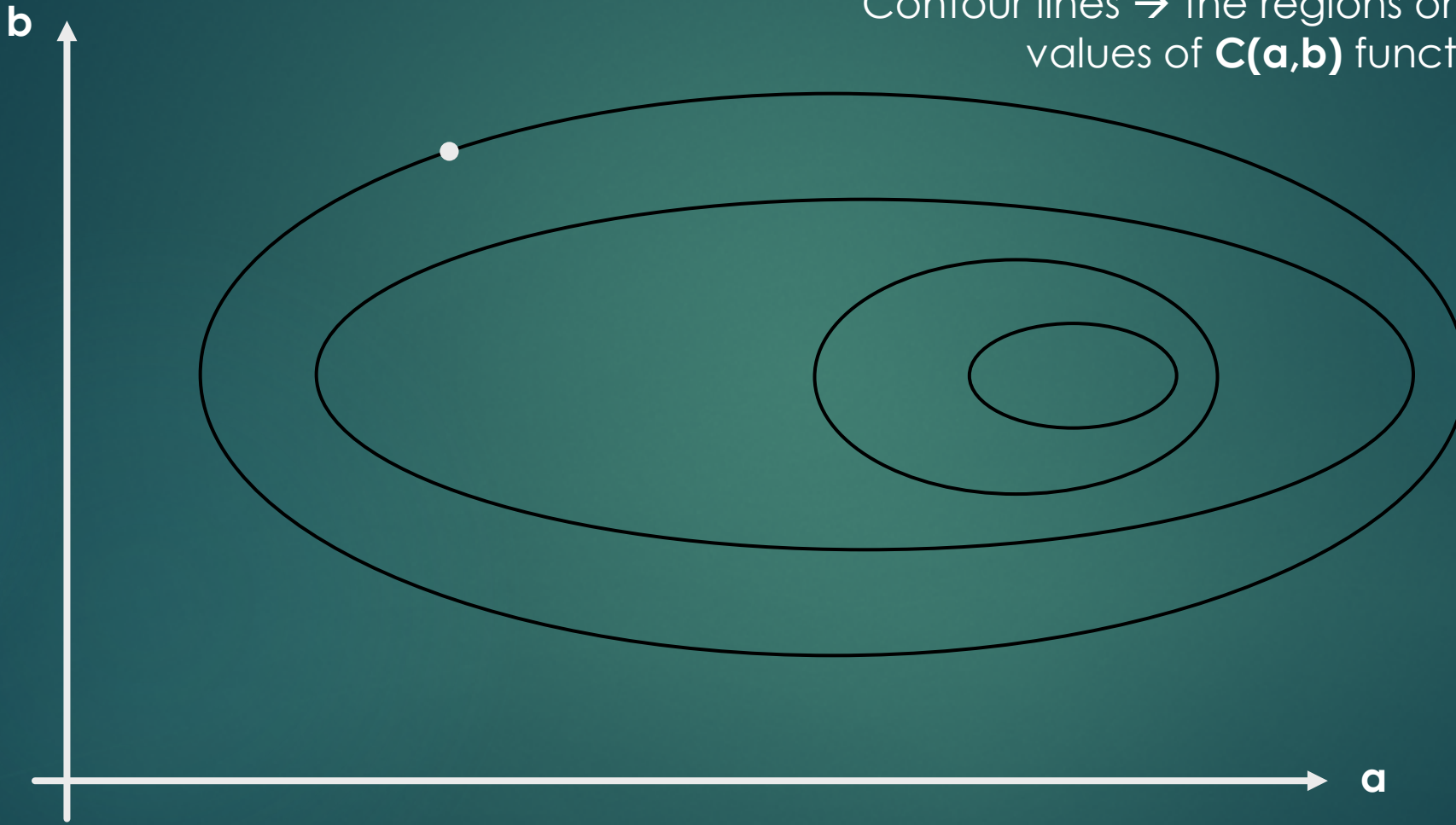
We have to know the partial derivative of the cost function
and go to the direction of the gradient
// gradient ~ partial derivative

$$\frac{\partial C(a, b)}{\partial a}$$

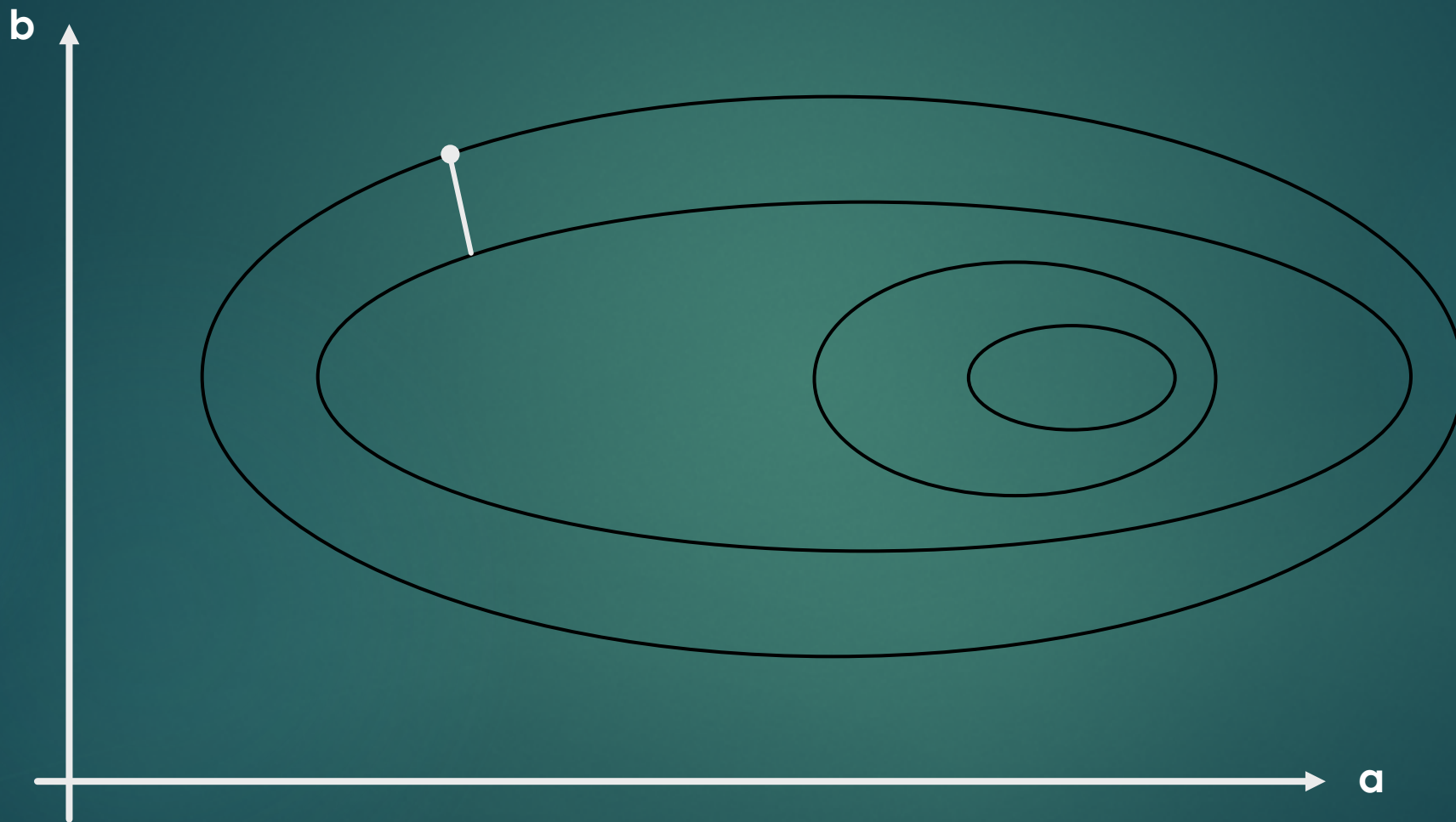
$$\frac{\partial C(a, b)}{\partial b}$$

Gradient descent

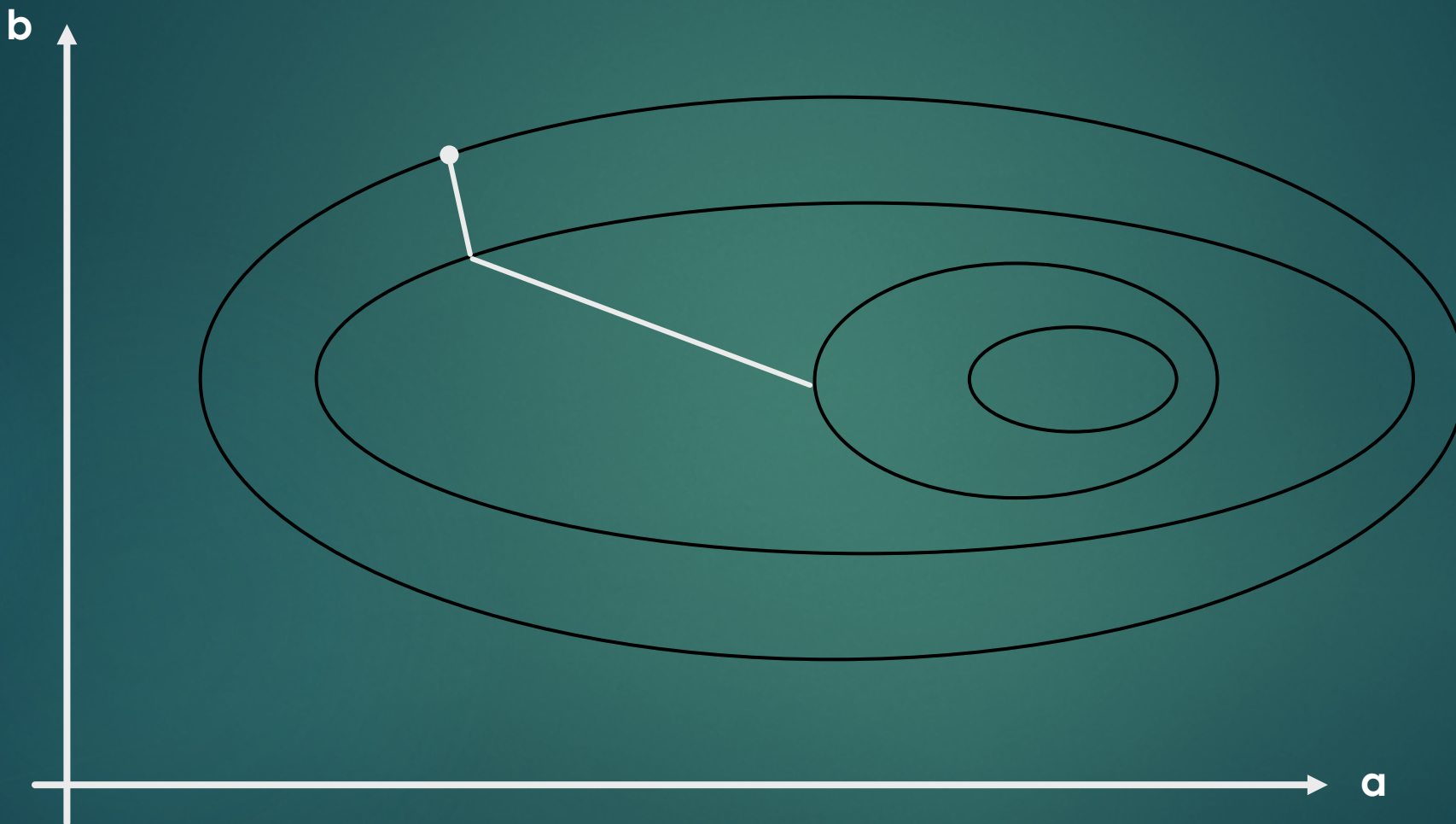
Contour lines \rightarrow the regions on which the values of $C(\mathbf{a}, \mathbf{b})$ function are constant !!!



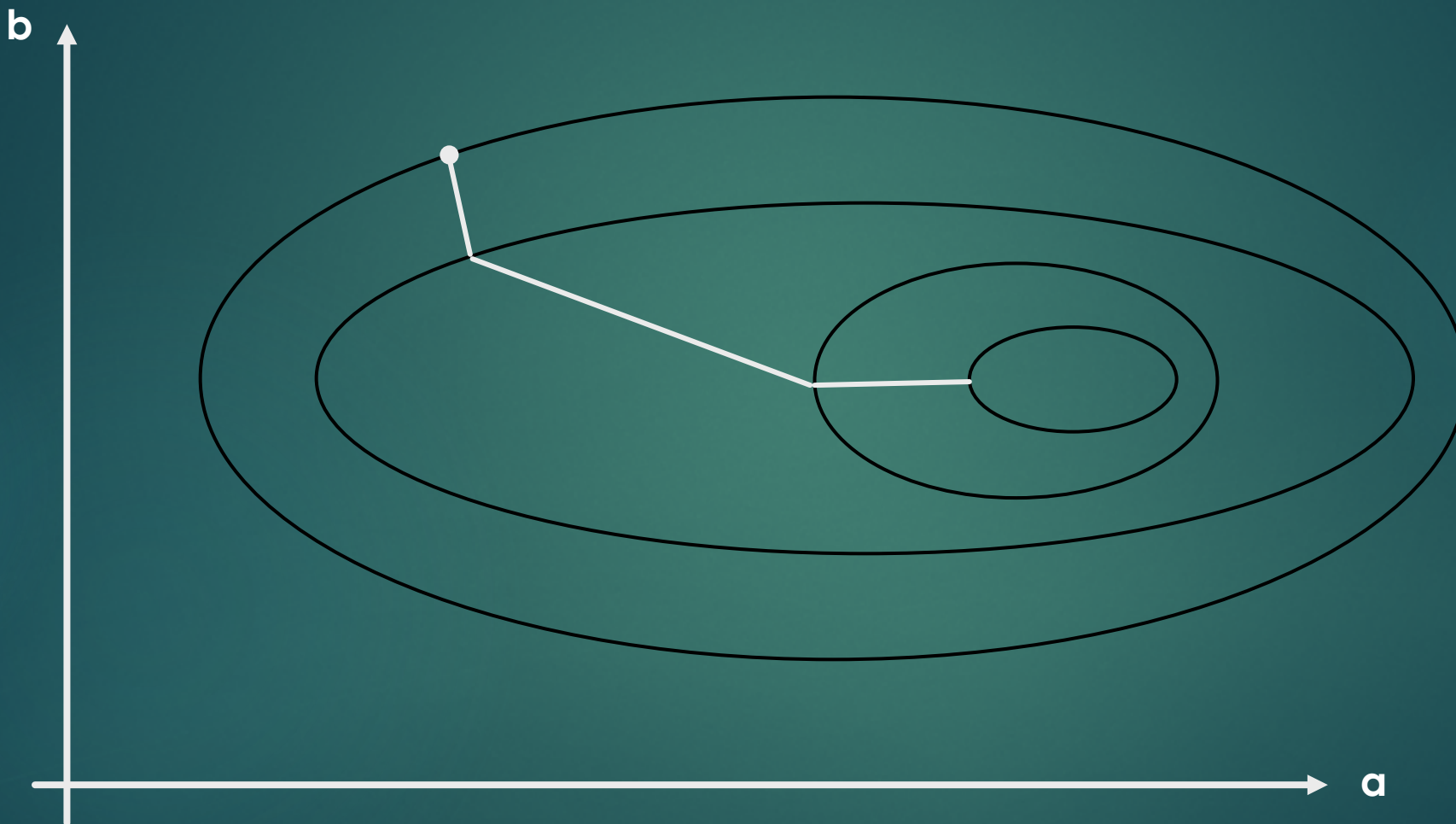
Gradient descent



Gradient descent



Gradient descent



Gradient descent

