Linear Regression

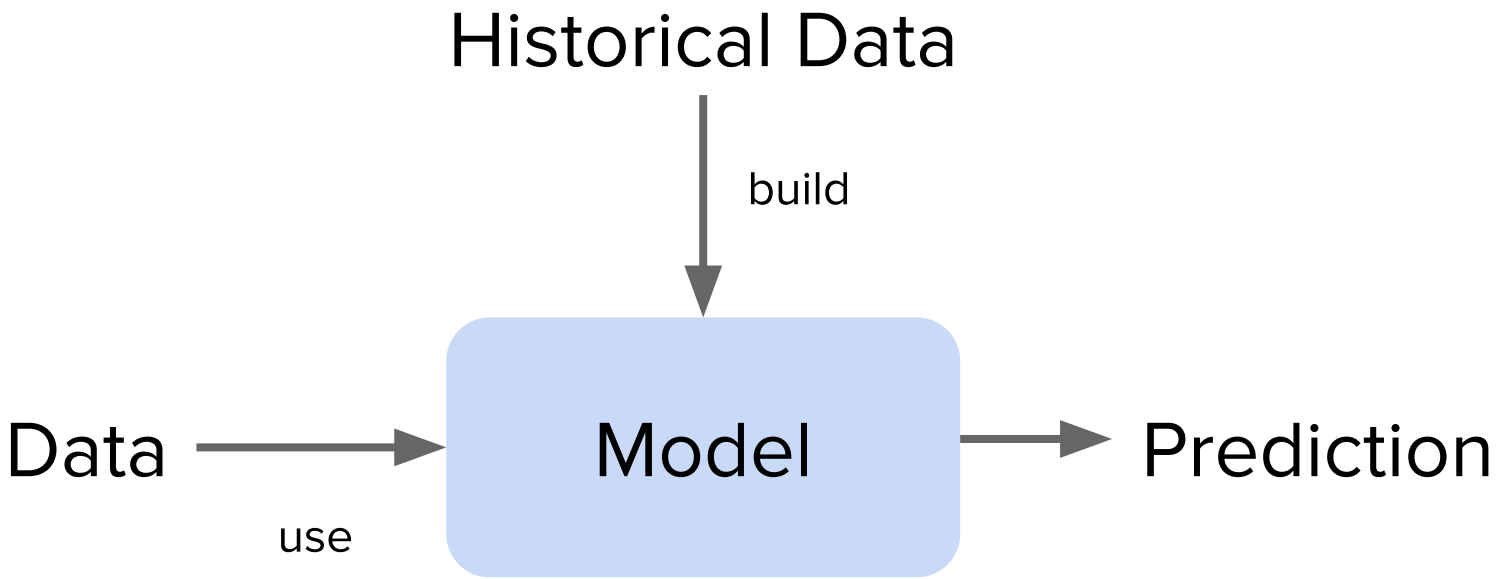
Week 04 - Day 01

Should I date this guy?

Congrats, you just used a model!

(Please appreciate how cool this gif is! :P)

What's a model?


Historical Data
build
Data
use
Model
Prediction

How long will it take to go to the airport?

Will Italy win against Germany?

(No, because Italy is not in the World Cup )

You use models every day

(transport, food, people, etc.)

“All models are wrong

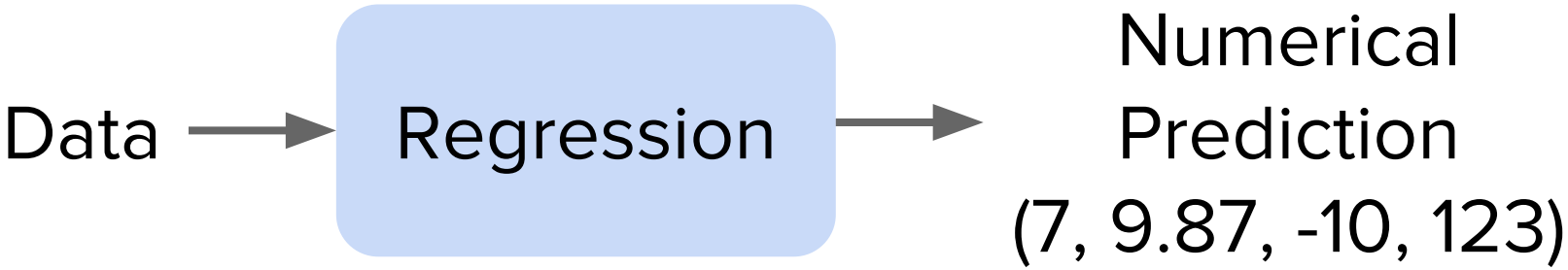
but some are useful”

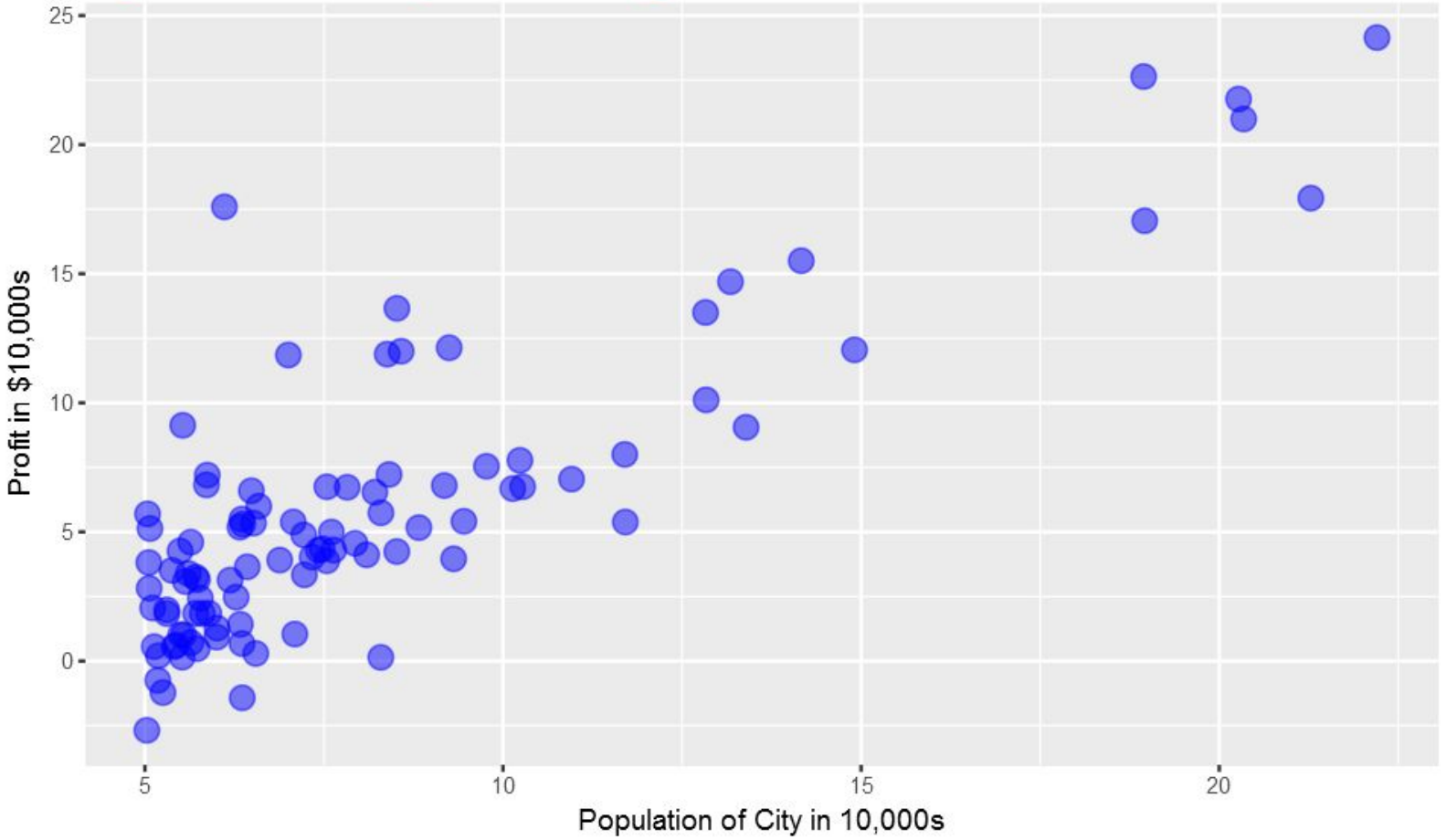
Regression

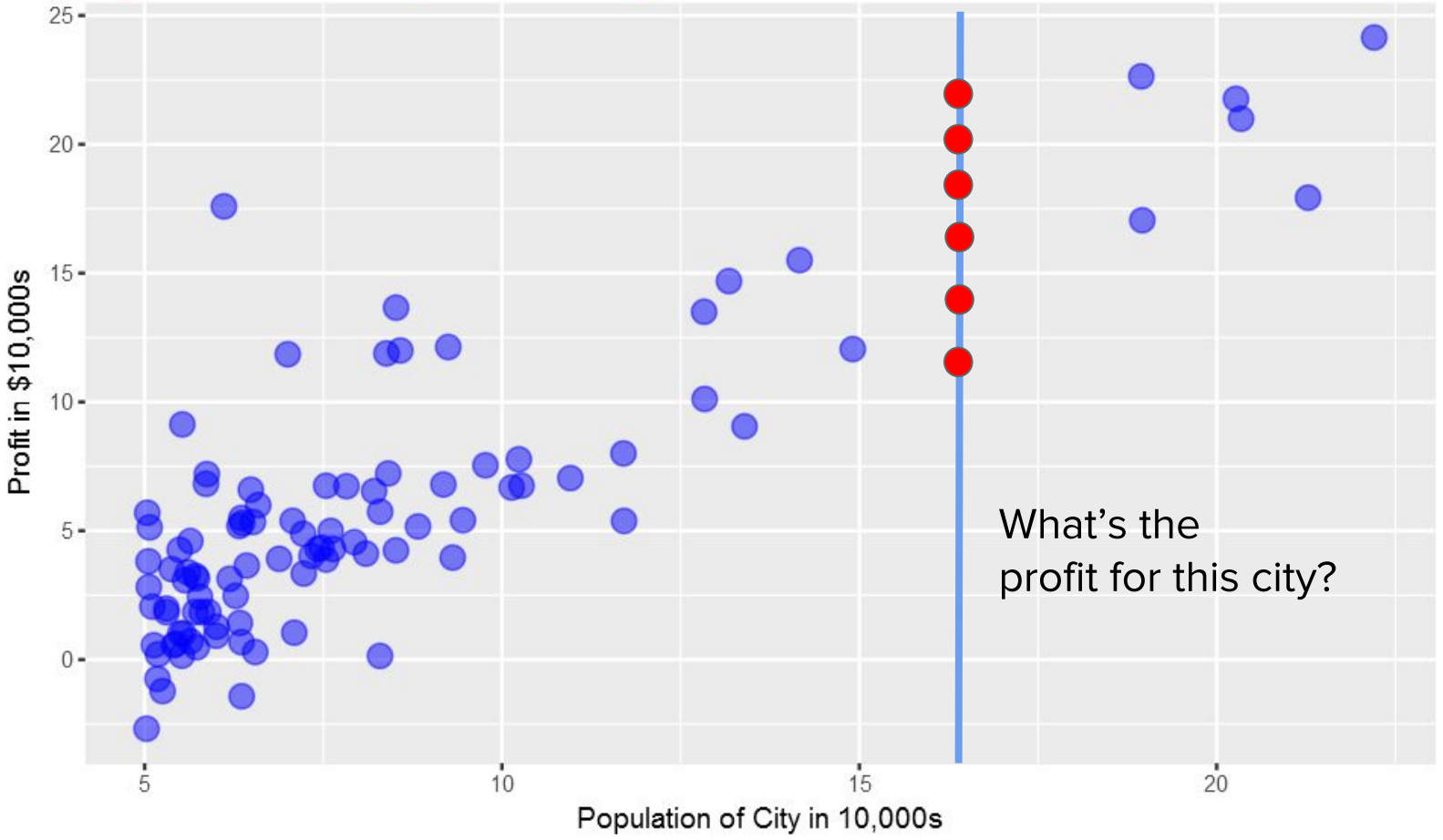
Regression

=

model for numerical values


Data
Regression
Numerical
Prediction
(7, 9.87, -10, 123)


Profit in $10,000s
25
20
15
10
5
0
5
10
15
20
Population of City in 10,000s

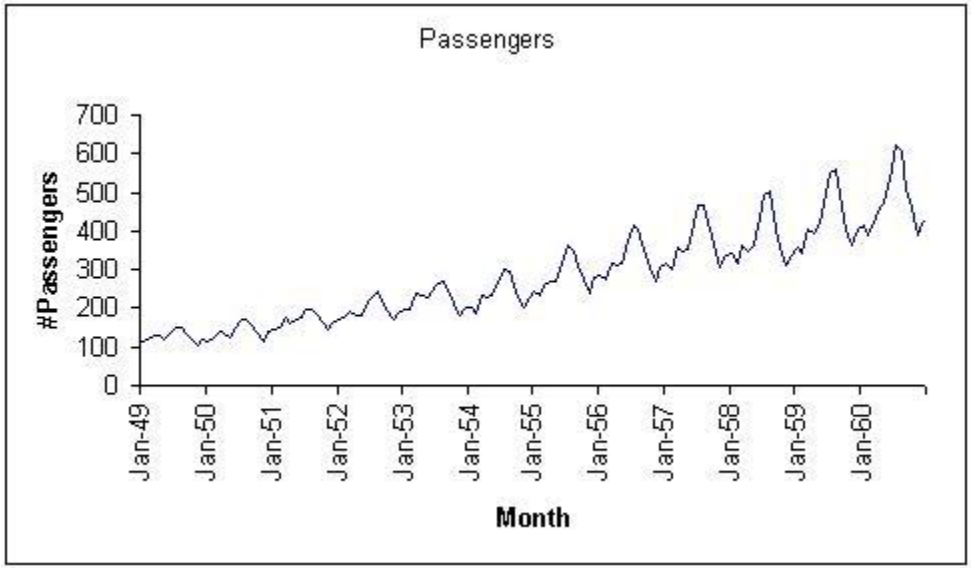

25
20
15
10
5
0
Profit in $10,000s
5
10
15
20
Population of City in 10,000s
What's the
profit for this city?

How many hours

will I need to complete project 3?

What will my salary be in 2 years?

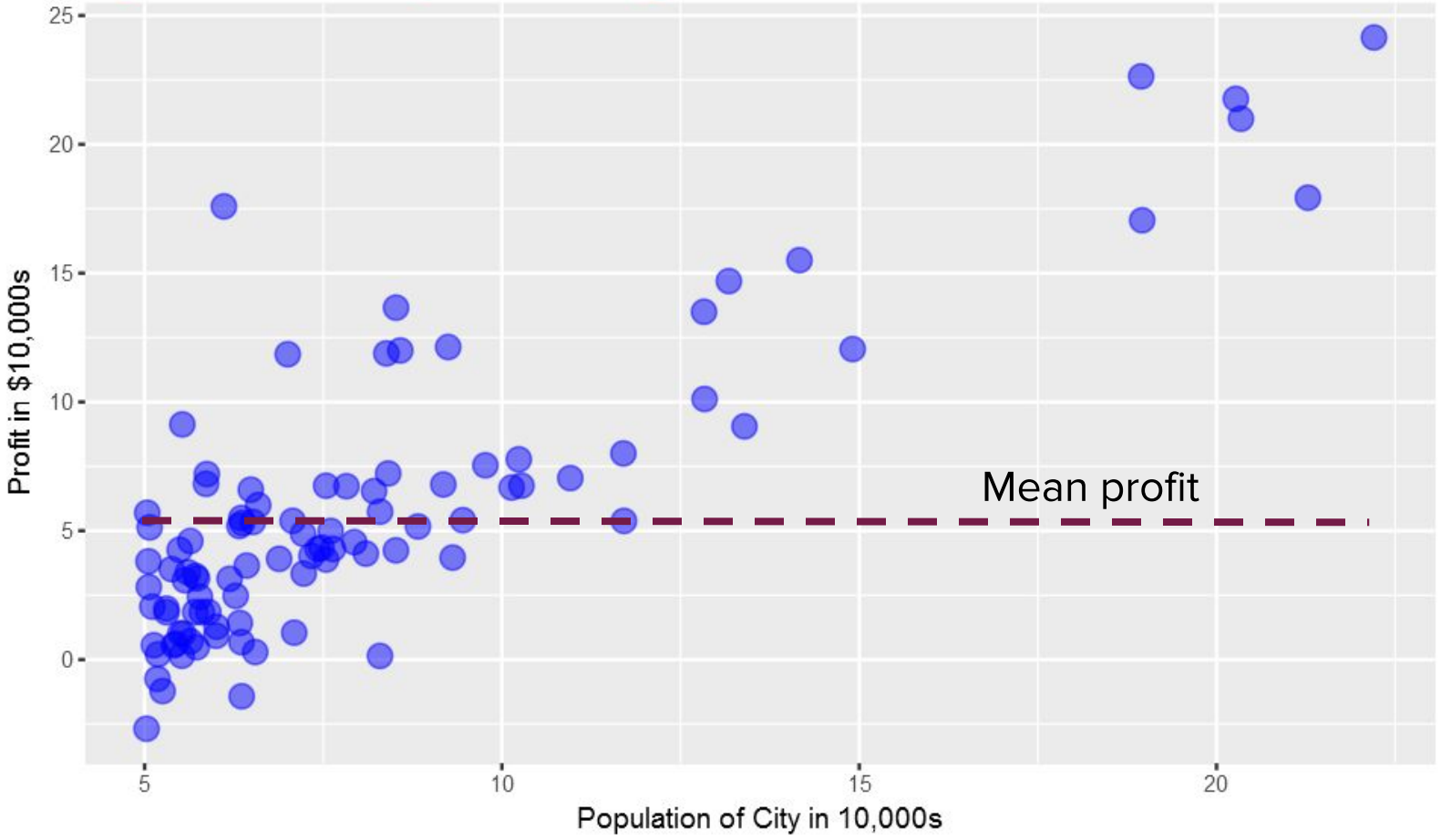
Regression != Time series analysis



Exercise

(notebook)

Dumb Regression (AKA Baseline)


Profit in $10,000s
25
20
15
10
5
0
5
10
15
20
Population of City in 10,000s
Mean profit

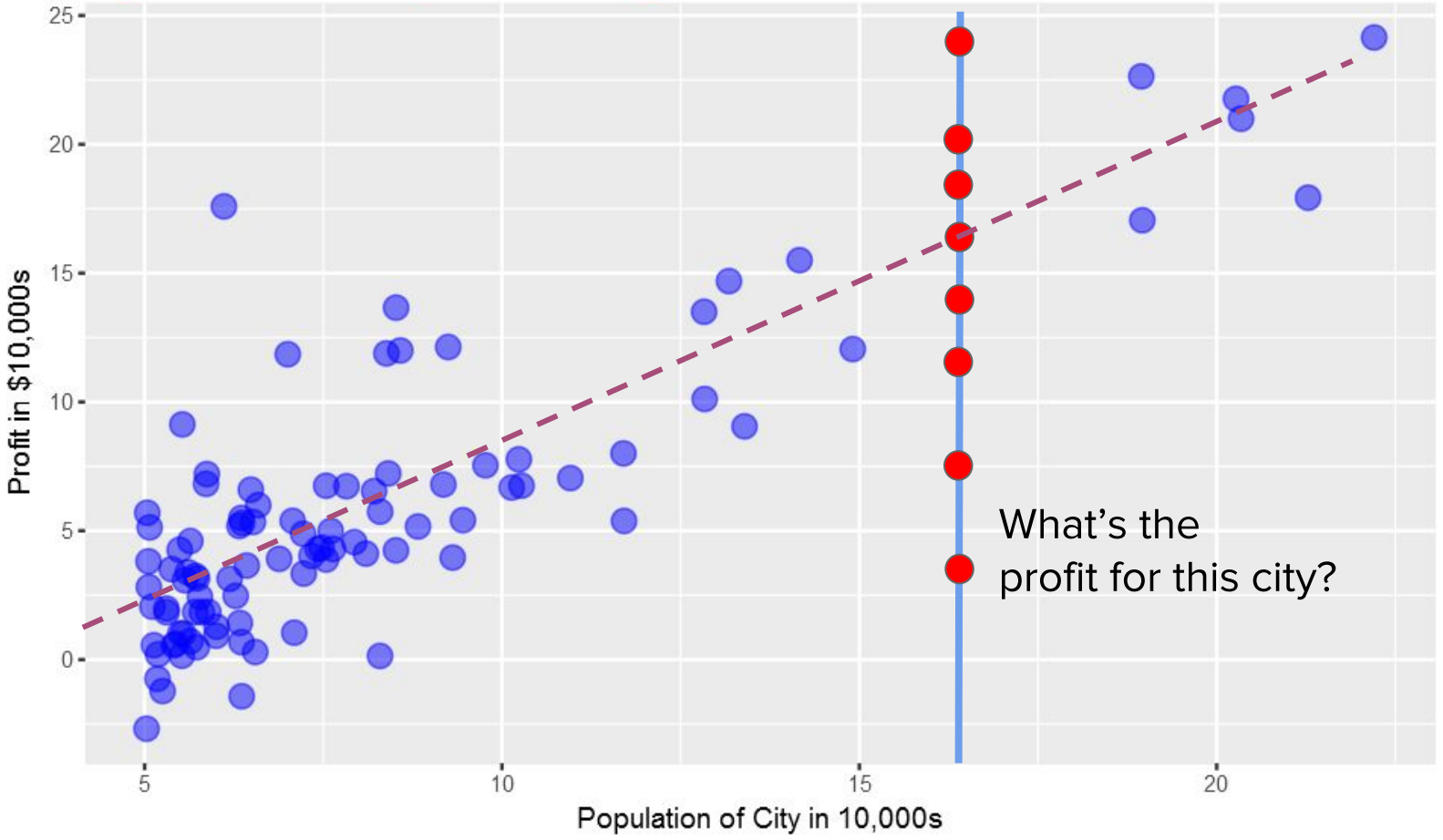
Linear regression

Linear regression

=

Linear relationship

Simple case = One predictor

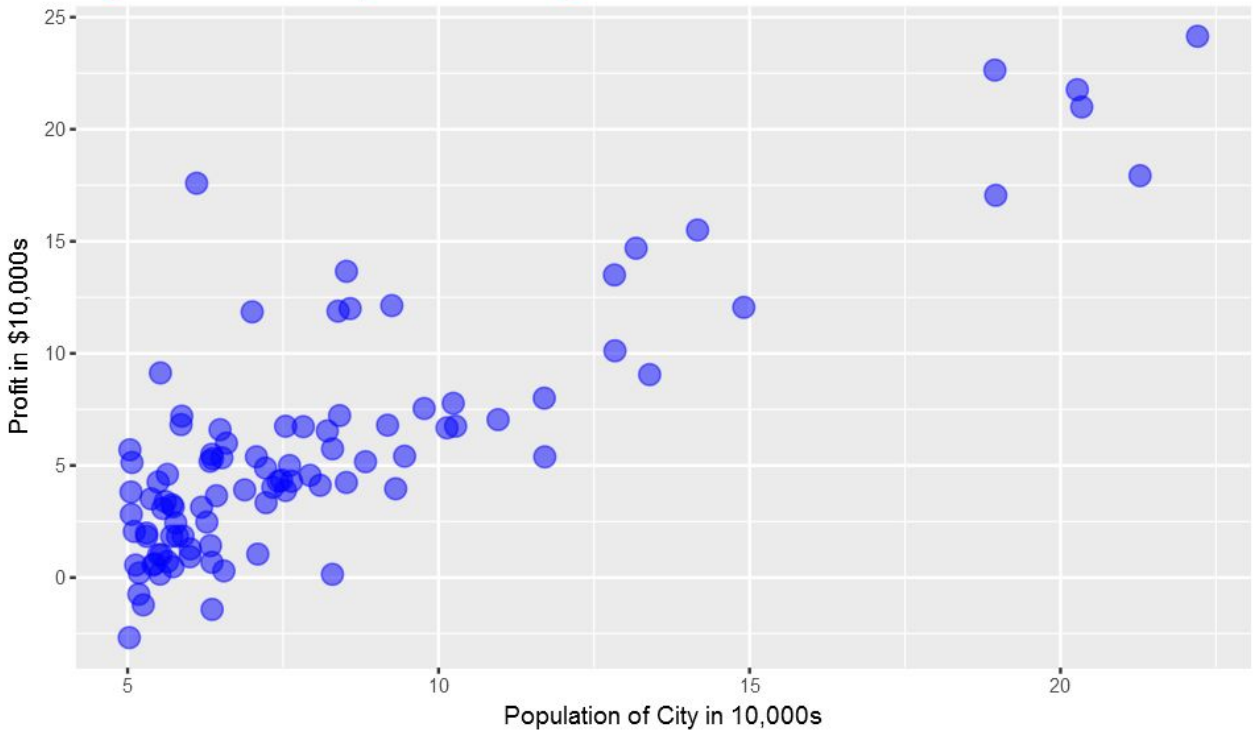

25
20
15
10
5
0
Profit in $10,000s
5
10
15
20
Population of City in 10,000s
What's the
profit for this city?

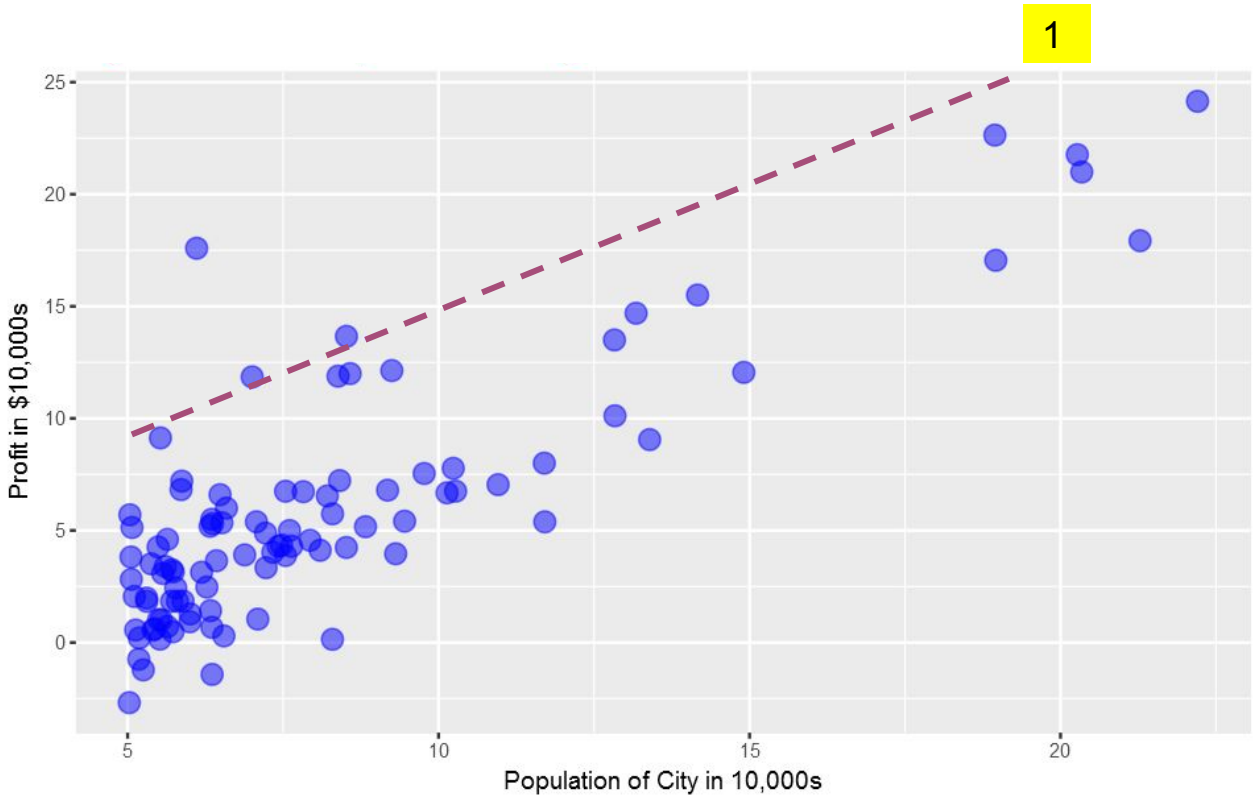
Slope + Intercept

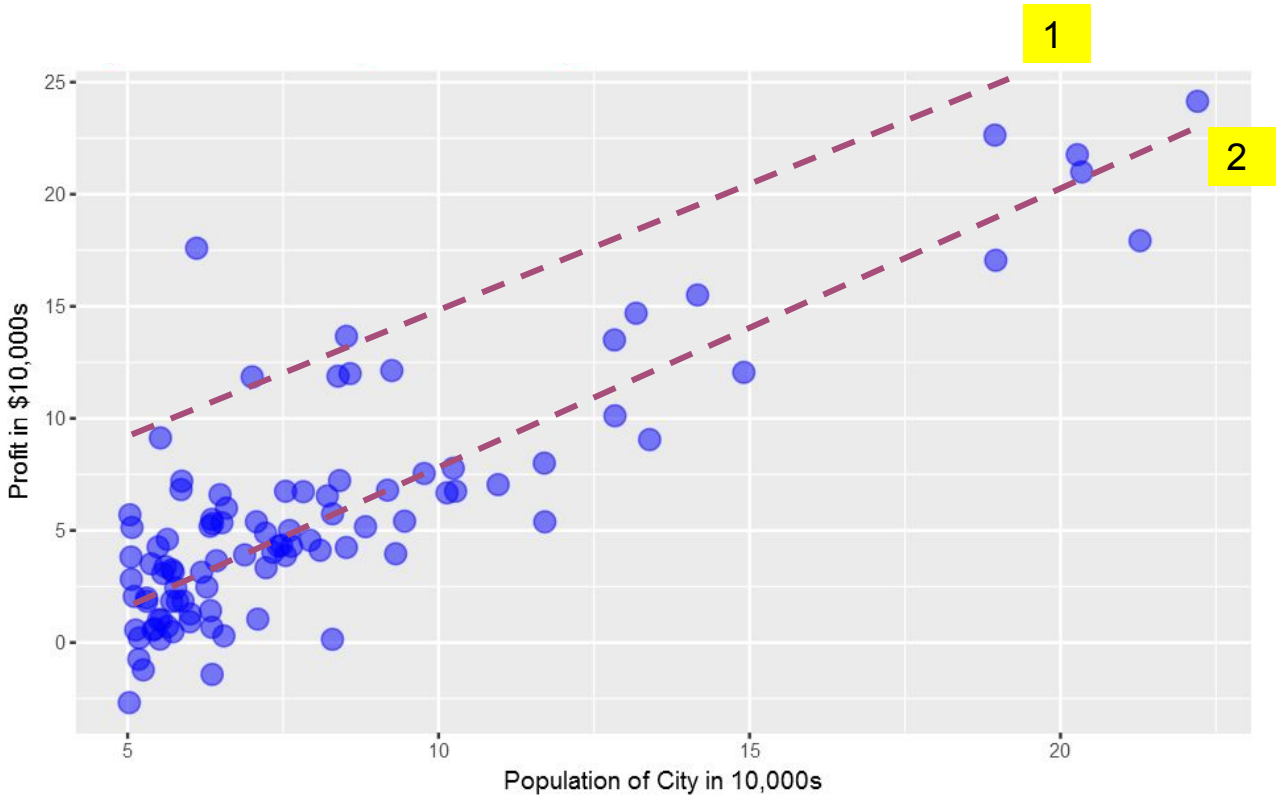
(parameters to define a line)

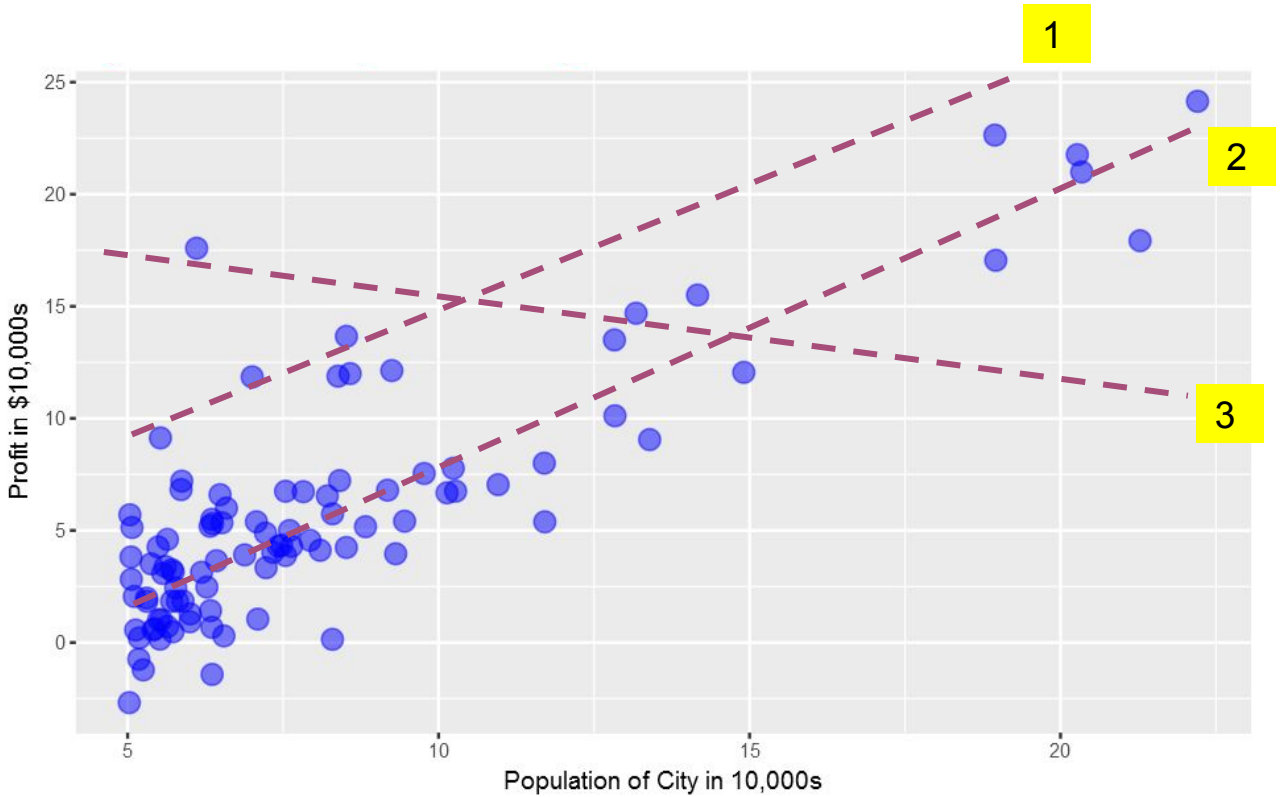
$$arrival\_time = 3 * mrt\_stops + 10$$

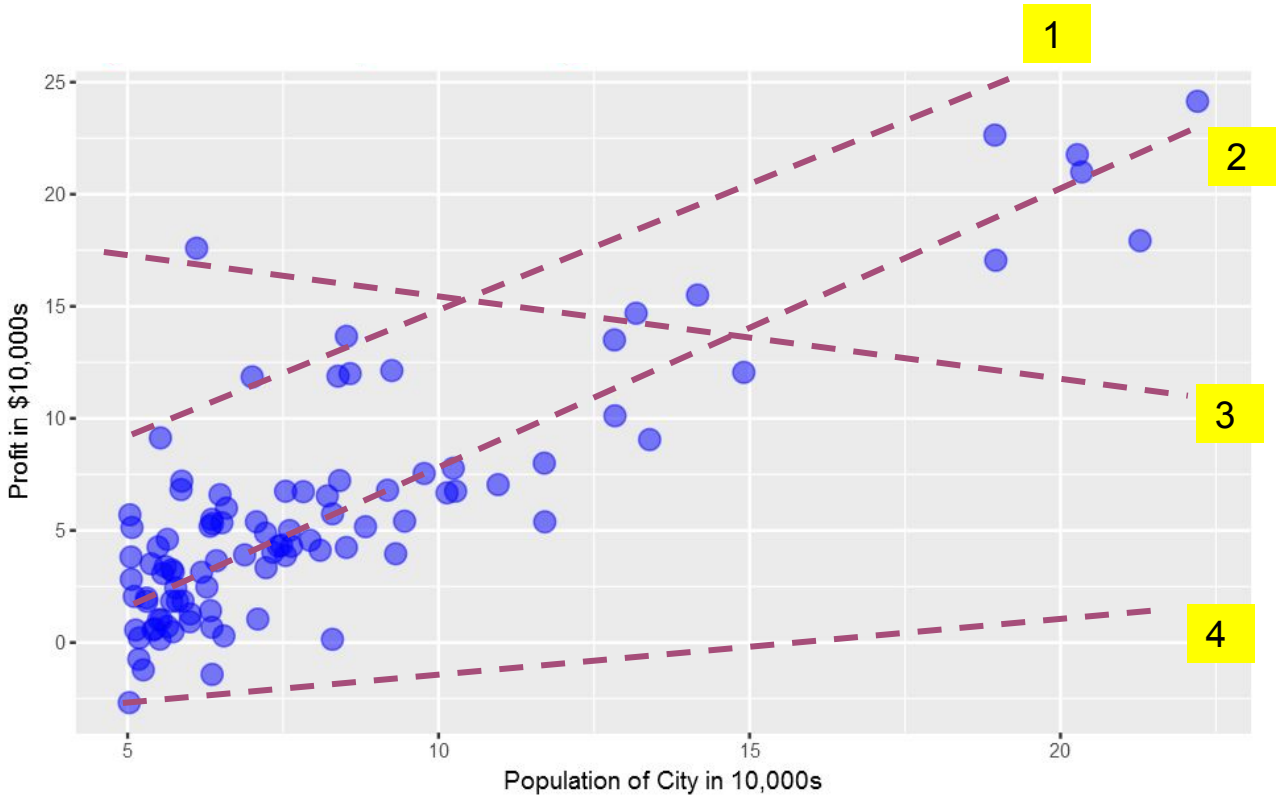
Which model is better? Why?


25
20
15
10
5
0
Profit in $10,000s
5
10
15
20
Population of City in 10,000s


1
25
20
15
10
5
0
Profit in $10,000s
5
10
15
20
Population of City in 10,000s


1
2
25
20
15
10
5
0
Profit in $10,000s
5
10
15
20
Population of City in 10,000s


1
2
3
25
20
15
10
5
0
Profit in $10,000s
5
10
15
20
Population of City in 10,000s


1
2
3
4
25
20
15
10
5
0
Profit in $10,000s
5
10
15
20
Population of City in 10,000s

Which one is better? Why?

Minimize the Error

Best model

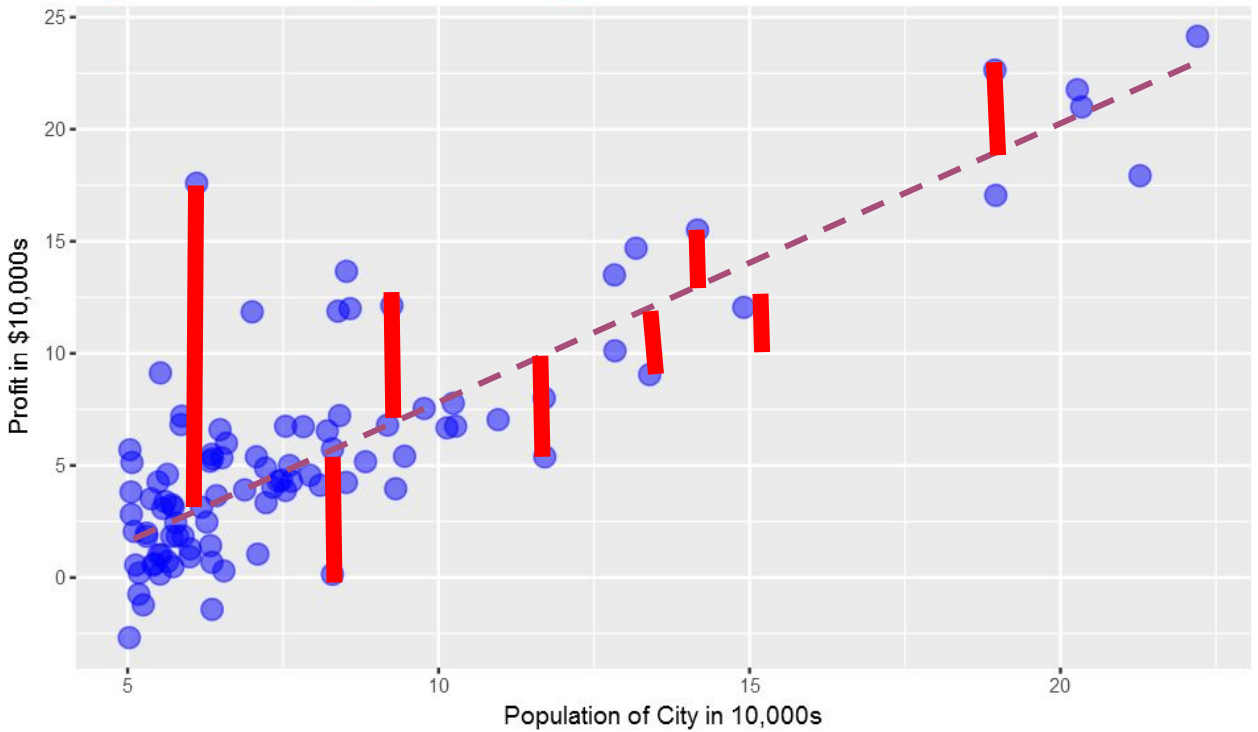
=

Smallest error

Formula for error?

Residual = real value - prediction

Residuals


25
20
15
10
5
0
Profit in $10,000s
5
10
15
20
Population of City in 10,000s

Sum of residuals

Absolute sum of residuals

Squared sum of residuals

Average absolute residual

Average squared residual

Sum of residuals

Absolute sum of residuals

Squared sum of residuals

Average absolute residual

Average squared residual

RSS = residual sum of squares

MSE = mean square error

MSE = RSS/n

Brute force approach

Simple case

=

one predictor

(and one intercept)

$$\text{prediction} = b0 + b1 * \text{predictor}$$

(this is the formula of a line!)

What’s the best combination of (b0,b1)?

```
for b0 in range(-100,100)
    for b1 in range (-100,100)
        print( b0, b1, get_error(data, b0, b1) )
```

-100, -100, 9469

-100,-99, 9321

...

100,99, 102934

100,100, 103563

Assumptions

Linearity

Independence

Normality

Equality of Variances

Linearity

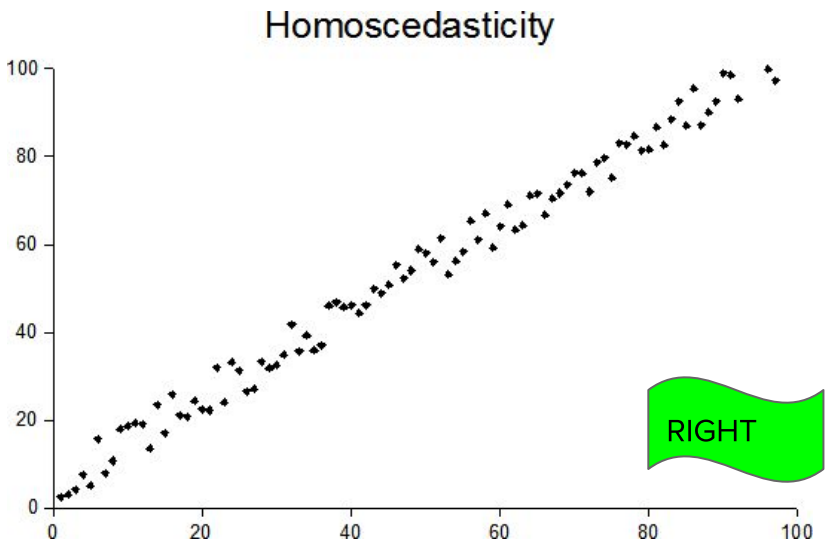
X and Y should have a linear relationship

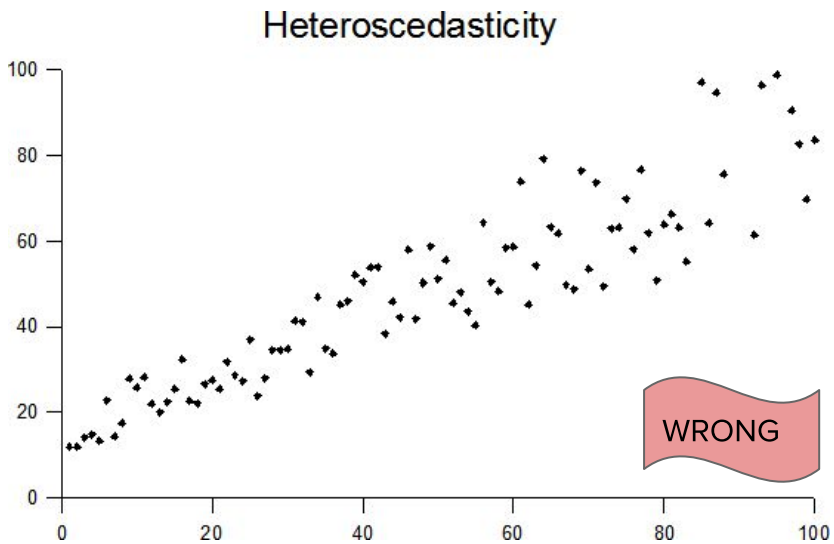
(remember the linear correlation?)

Independence

The residuals should be independent one from each other

Equality of Variances





Liner
....in the coefficients!

$$z = \log(x)$$

$$y = b0 + b1 * z$$

$$z = x^2$$

$$y = b0 + b1 * z$$

Categorical Features

$$\text{Salary} = 50{,}000 - 5{,}000 * \text{is\_female}$$

Finding the intercept with Pearson's correlation coefficient

Only for the lesson/lab!

RECAP

1. Models are used to predict/classify
2. Models are built with historical data
3. We use models every day
4. Regression = predicting numbers
5. Linear regression
6. Best solution = smallest error