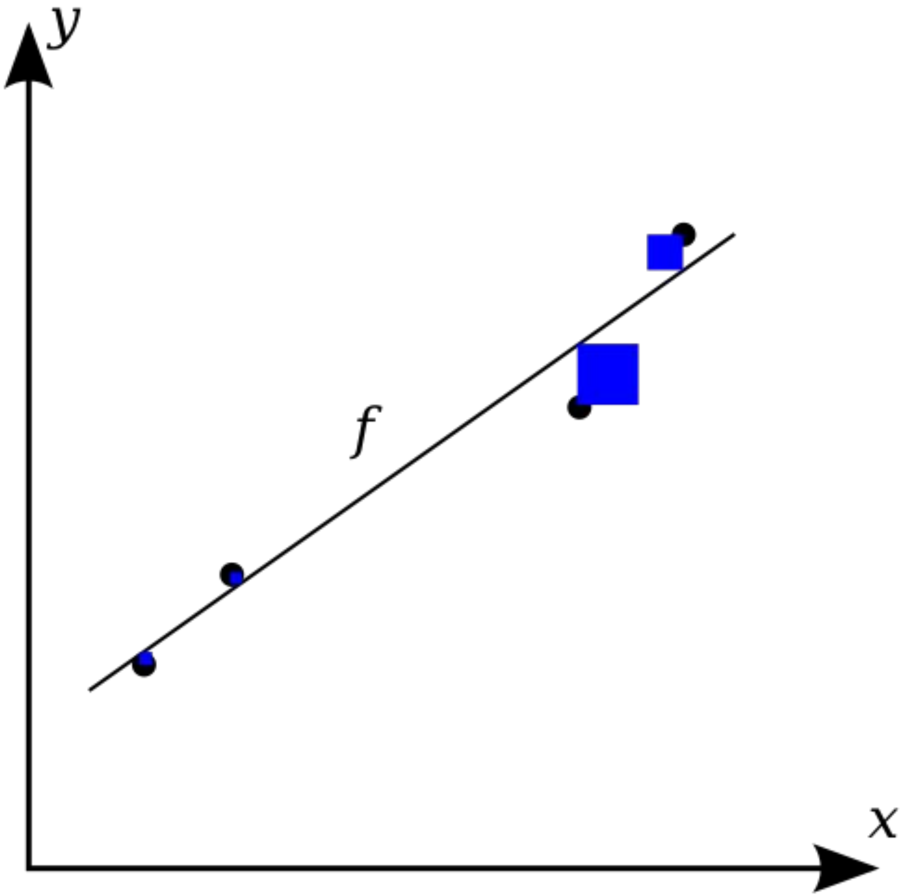
R2, MLR

Week 04 - Day 02

From yesterday

Error = sum of squared residuals = RSS



RSS = Blue area

Mean of y = a dumb baseline

Baseline error = 7

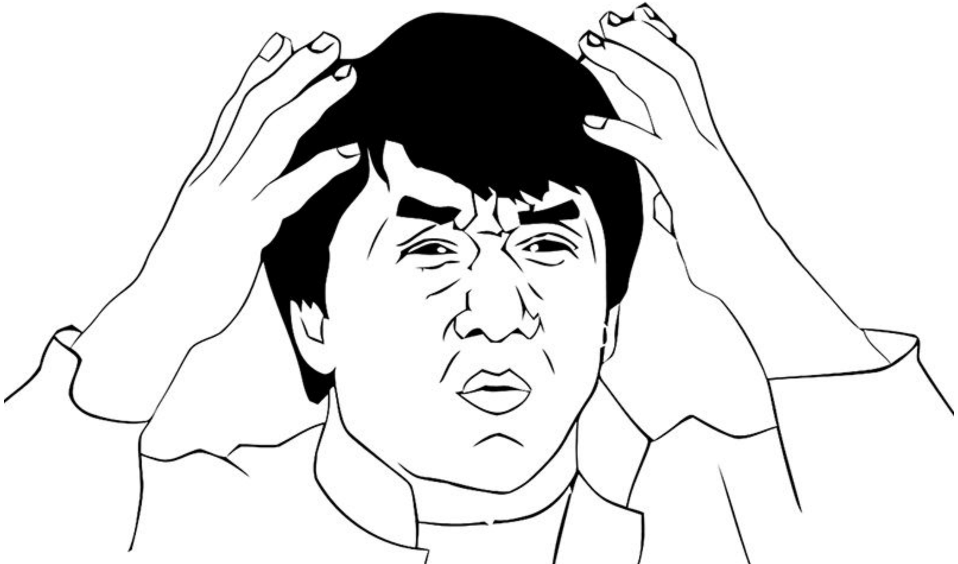
Model error = 6.9

Is it a good model?

R Squared

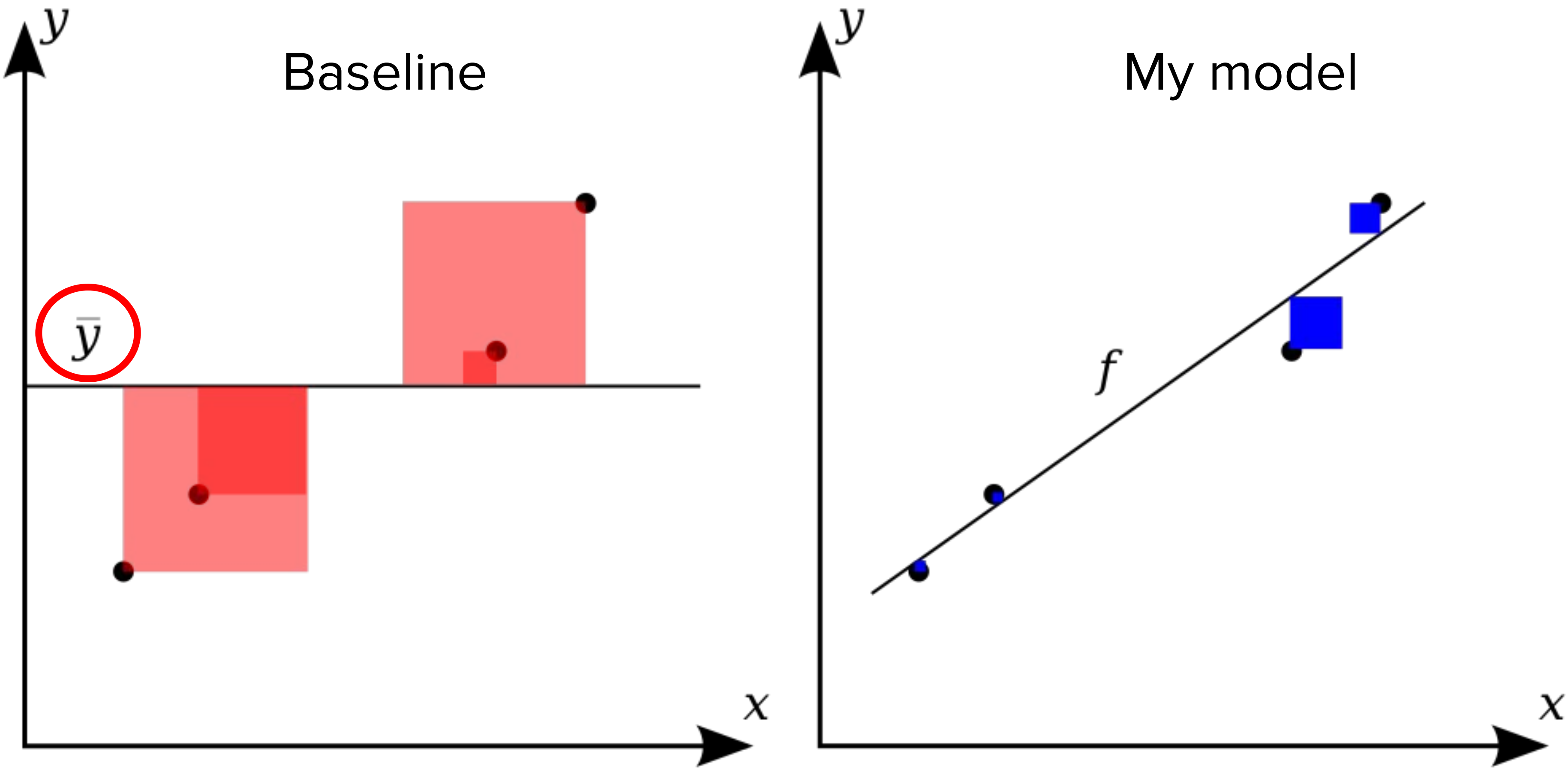
"R2 is the proportion of the variance in the dependent variable that is predictable from the independent variable(s)."

Wikipedia



Intuition

Is my model better than the baseline?


Baseline
y
x
ȳ
My model
y
x
f

blue_area == red_area

Meaning?

blue_area < red_area

Meaning?

blue_area > red_area

Meaning?

R2 = 1 - (blue_area / red_area)

Example 1 (good model)

My model error = 10

Baseline error = 100

$R2 = 1 - 10/100 = 1 - 0.1 = 0.9$

Example 2 (ok model)

My model error = 50

Baseline error = 100

R2 = 1 - 50/100 = 1 - 0.5 = 0.5

Example 3 (bad model)

My model error = 90

Baseline error = 100

R2 = 1 - 90/100 = 1 - 0.9 = 0.1

Is R2 enough to say a model is good or bad?

No, but it gives a good indication!

Optimization Using Derivative

Prediction = b0 + b1*input

Time = 10 + 3 * mrt_stops

Optimization

=

Finding b0,b1 that give

the smallest error

Simple example:

just one coefficient (slope)

no intercept

Random error we
cannot control/model

salary = b1 * year_experience + $\boxed{\text{error}}$

Residuals

$$\text{Error} = \text{SUM}((\overset{\text{Ground truth}}{\boxed{\text{y\_hat}}} - (\overset{\text{Prediction}}{\boxed{\text{b0+b1*x}}}))\text{**}2)$$

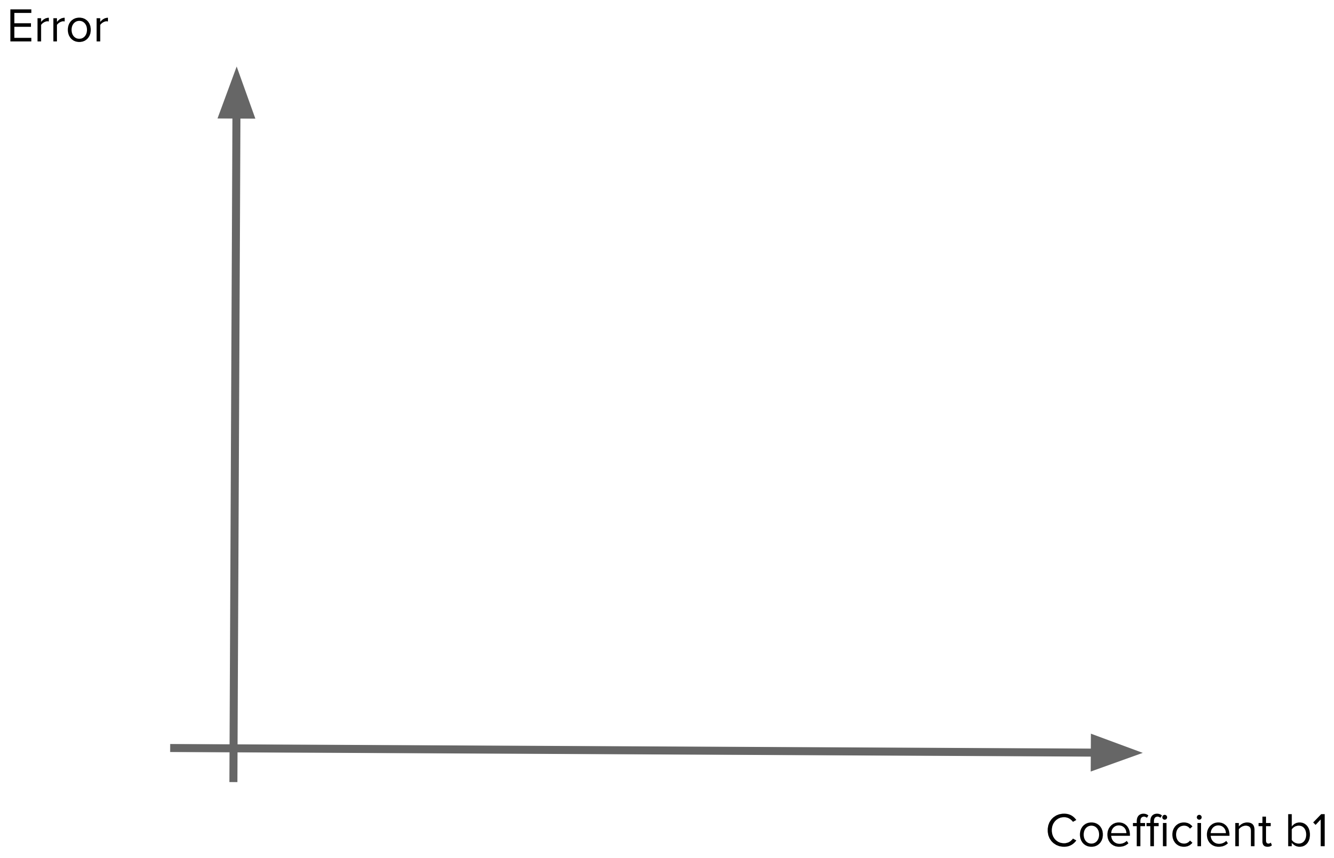
$$\text{Error} = \text{SUM}((\overset{\text{Ground truth}}{\boxed{\text{y\_hat}}} - (\overset{\text{Prediction}}{\boxed{\text{b0+b1*x}}}))\text{**}2)$$

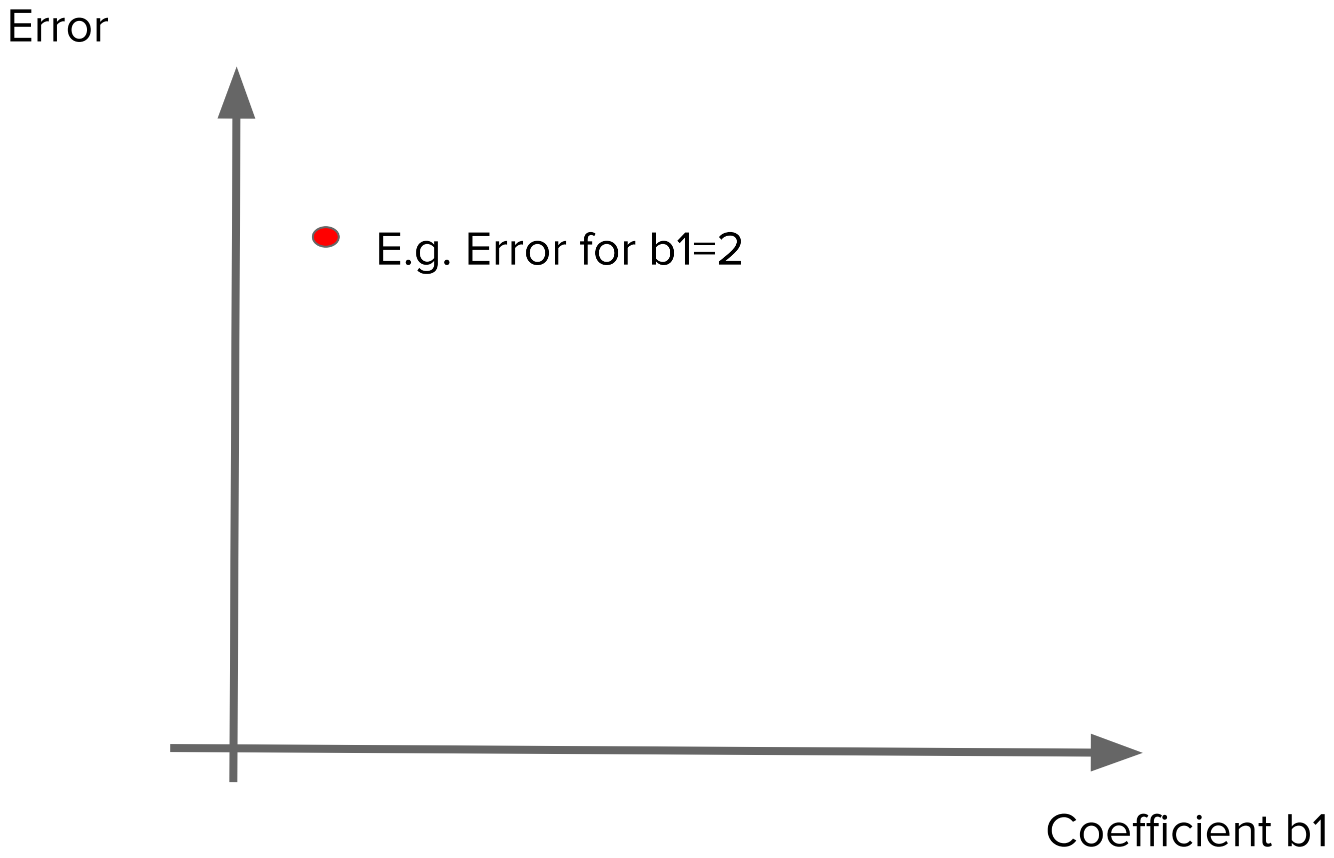
$$\text{Error} = \text{f(b0,b1)}$$

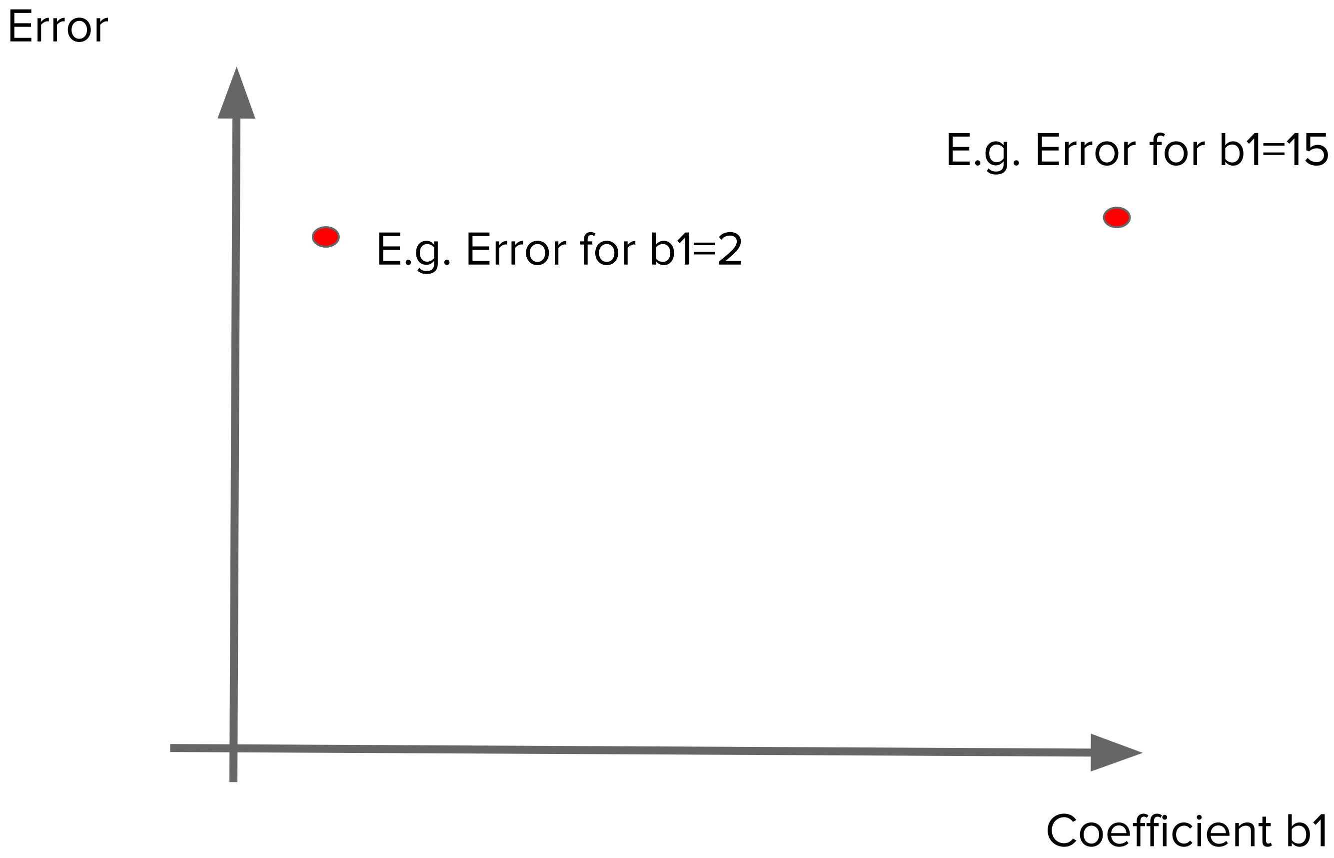
$$\text{Error} = \text{SUM}((\overset{\text{Ground truth}}{\boxed{\text{y\_hat}}} - (\overset{\text{Prediction}}{\boxed{\text{b0+b1*x}}}))\text{**2})$$

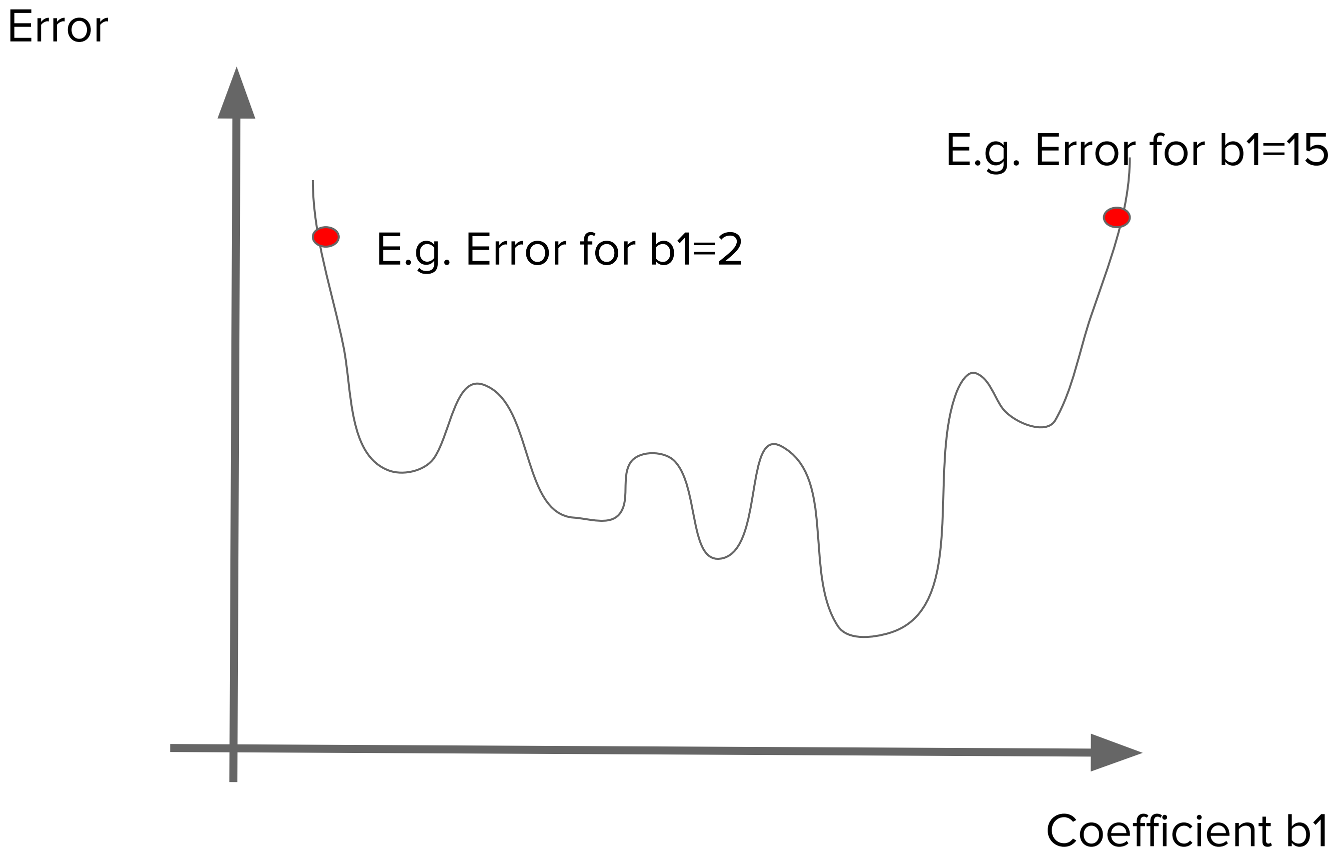
$$\text{Error} = \text{f}(\cancel{\text{b0}},\text{b1})$$

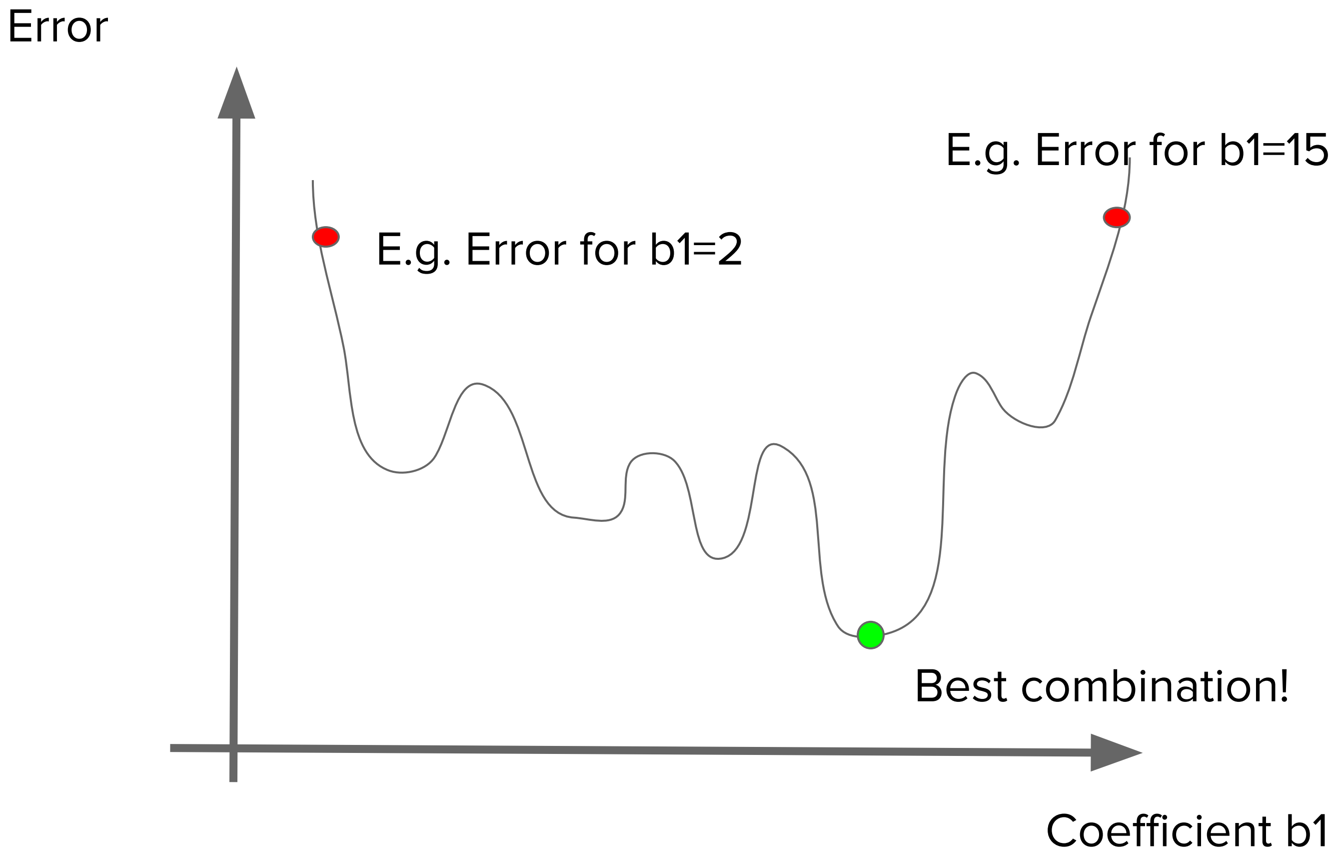
To keep it simple


Error
Coefficient b1

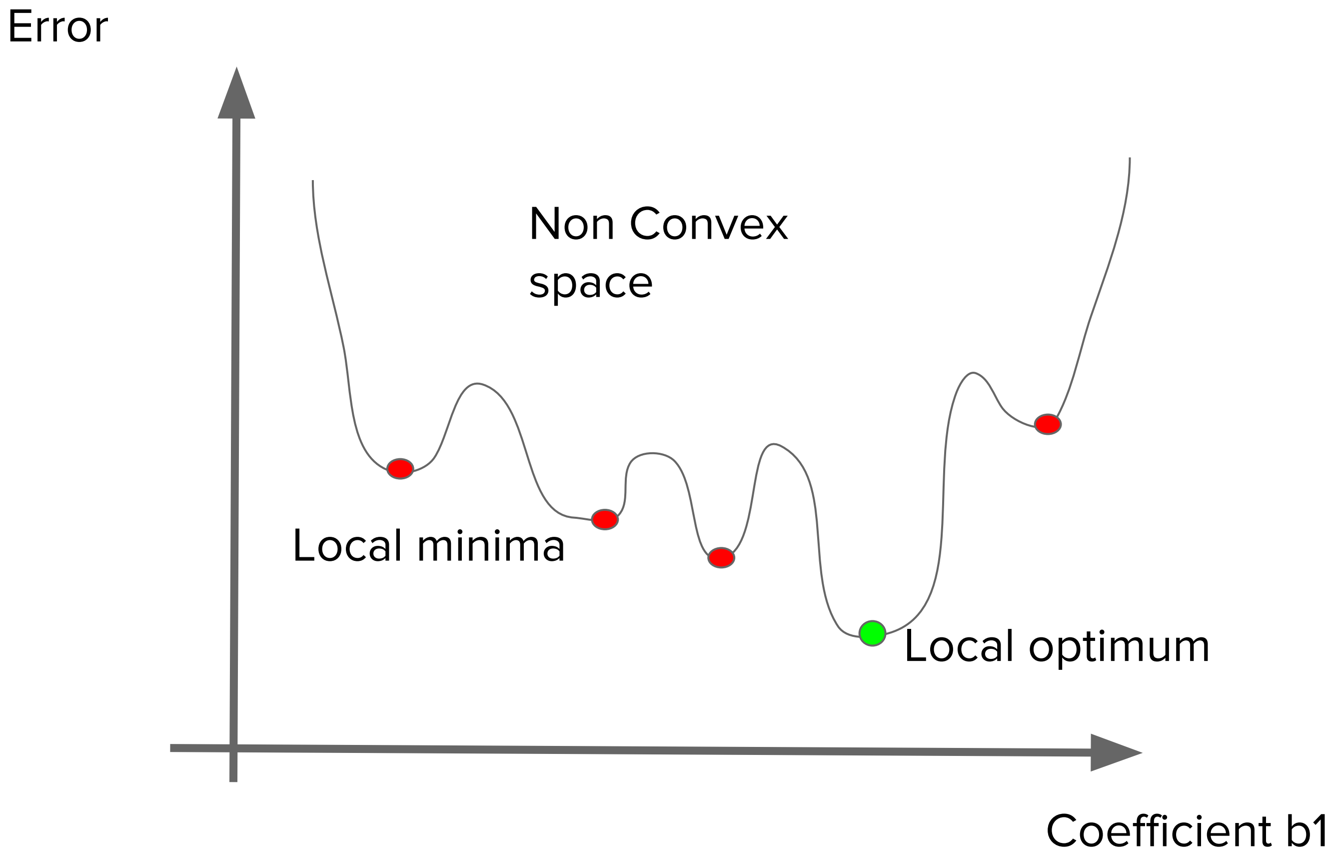

Error
E.g. Error for b1=2
Coefficient b1


Error
E.g. Error for b1=15
E.g. Error for b1=2
Coefficient b1

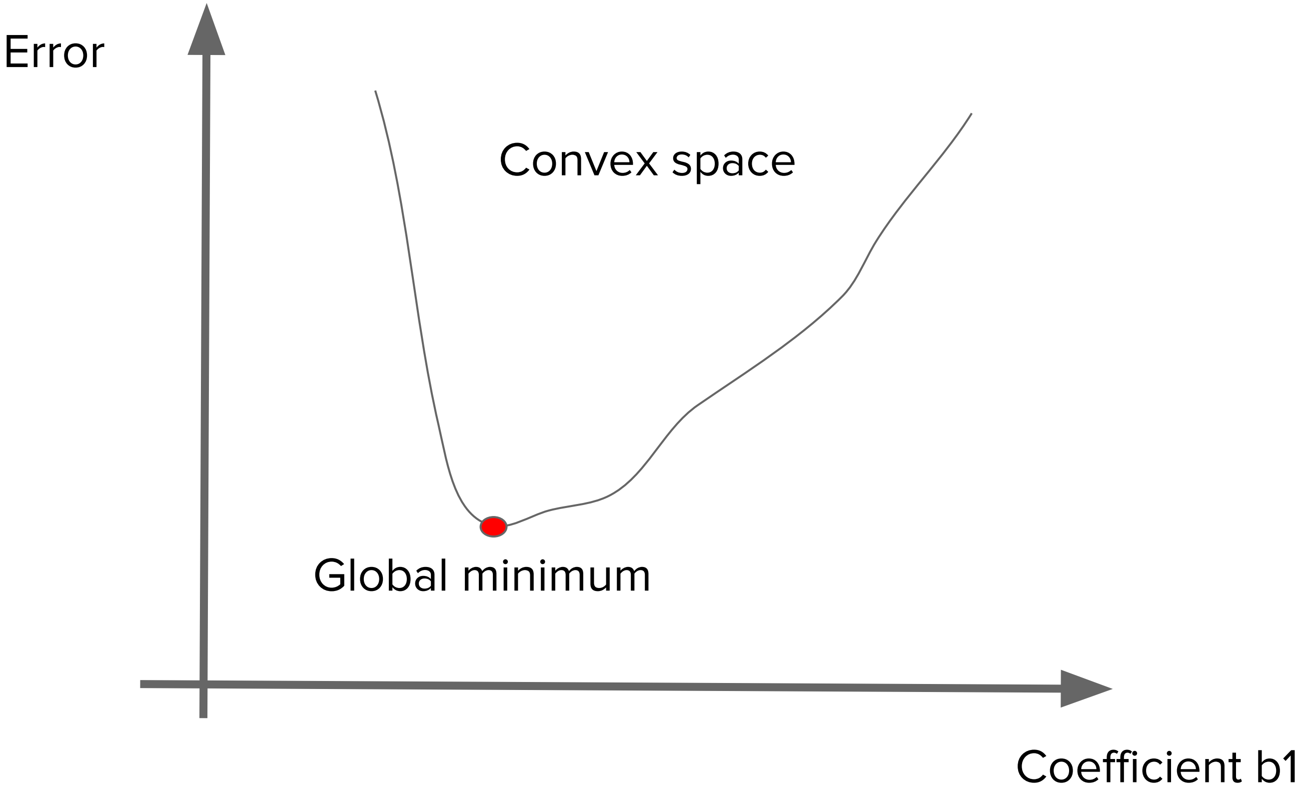

Error
E.g. Error for b1=15
E.g. Error for b1=2
Coefficient b1

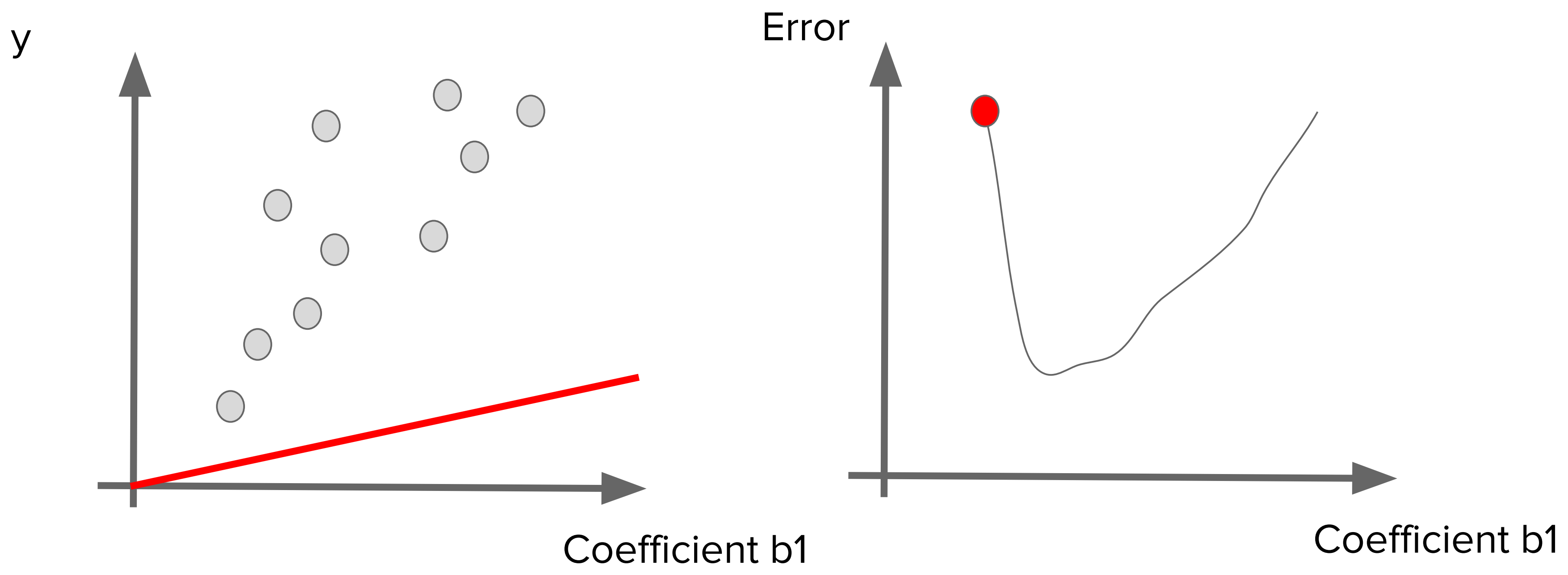


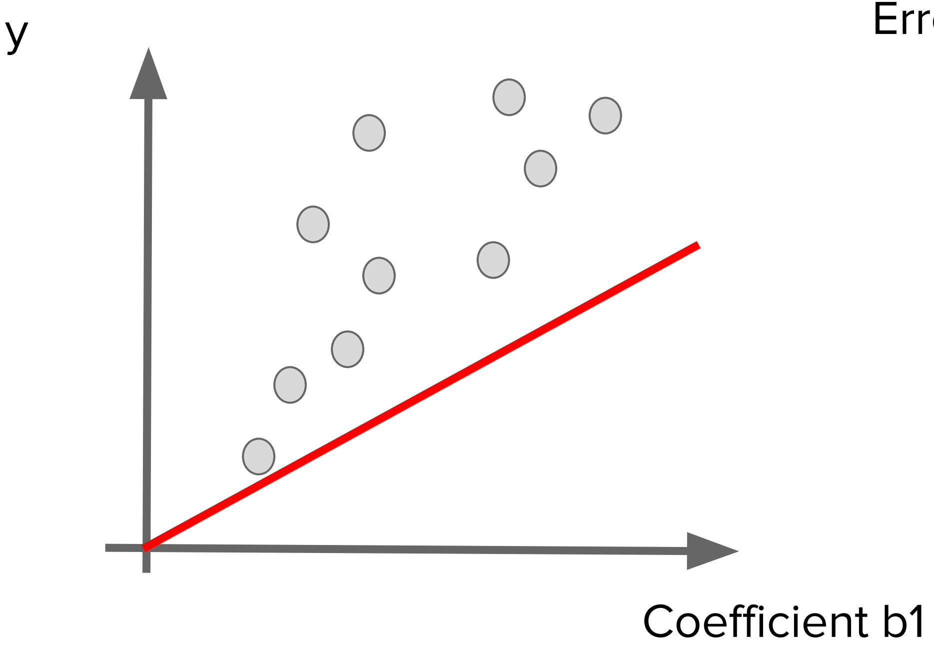
Error
E.g. Error for b1=15
E.g. Error for b1=2
Best combination!
Coefficient b1

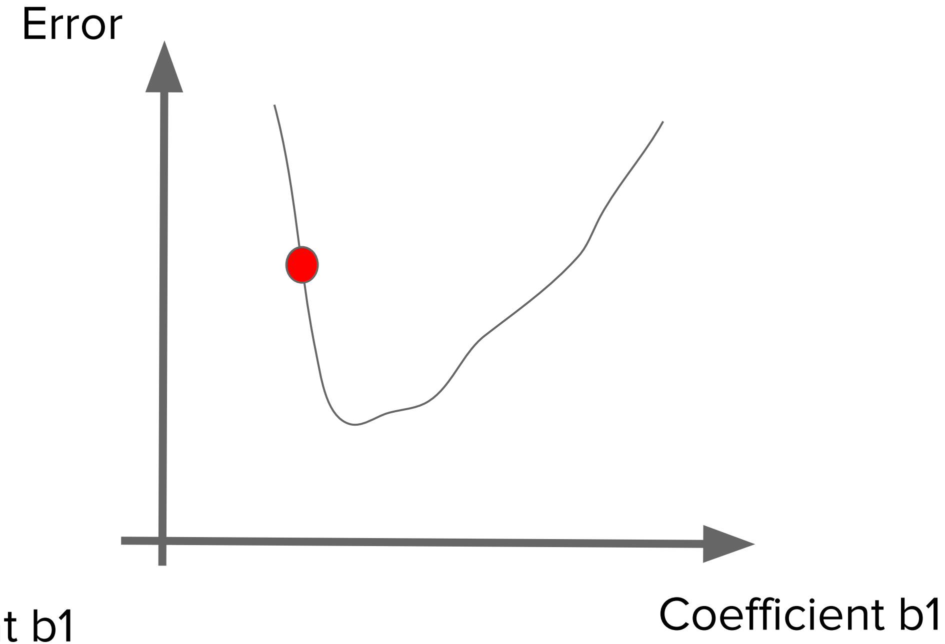

Error
Non Convex
space
Local minima
Local optimum
Coefficient b1

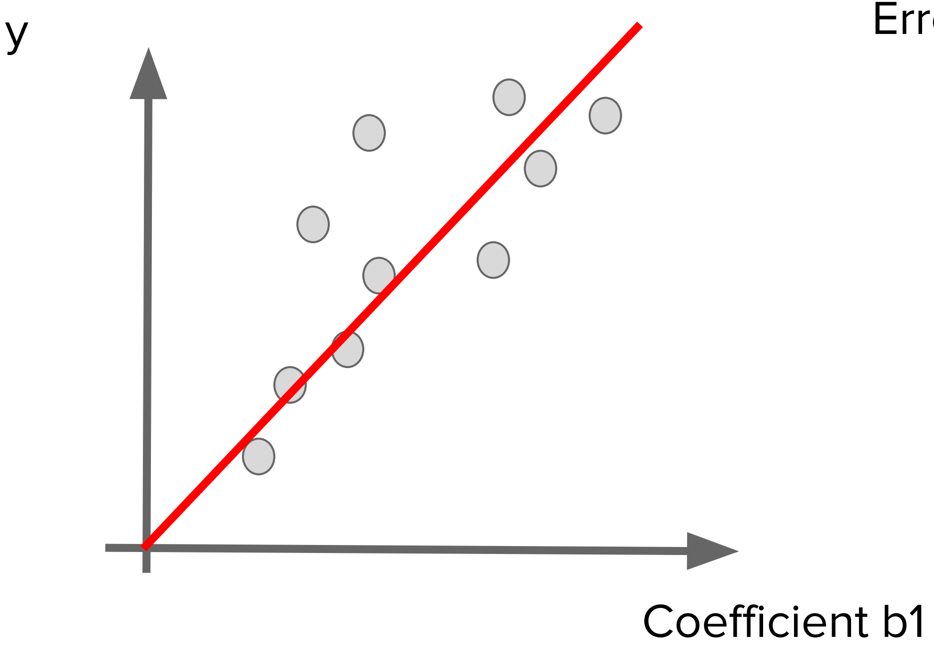
We are really lucky!

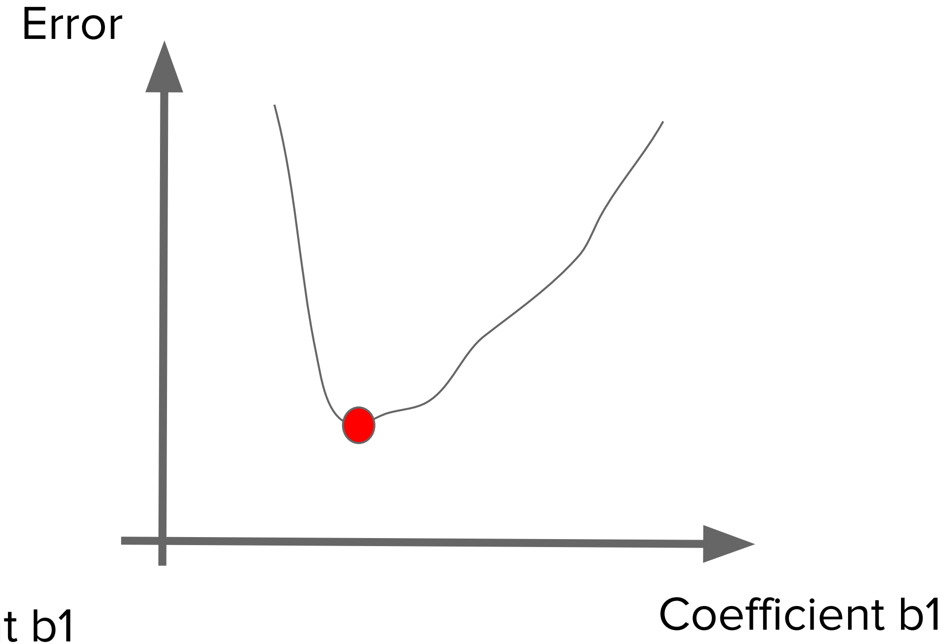


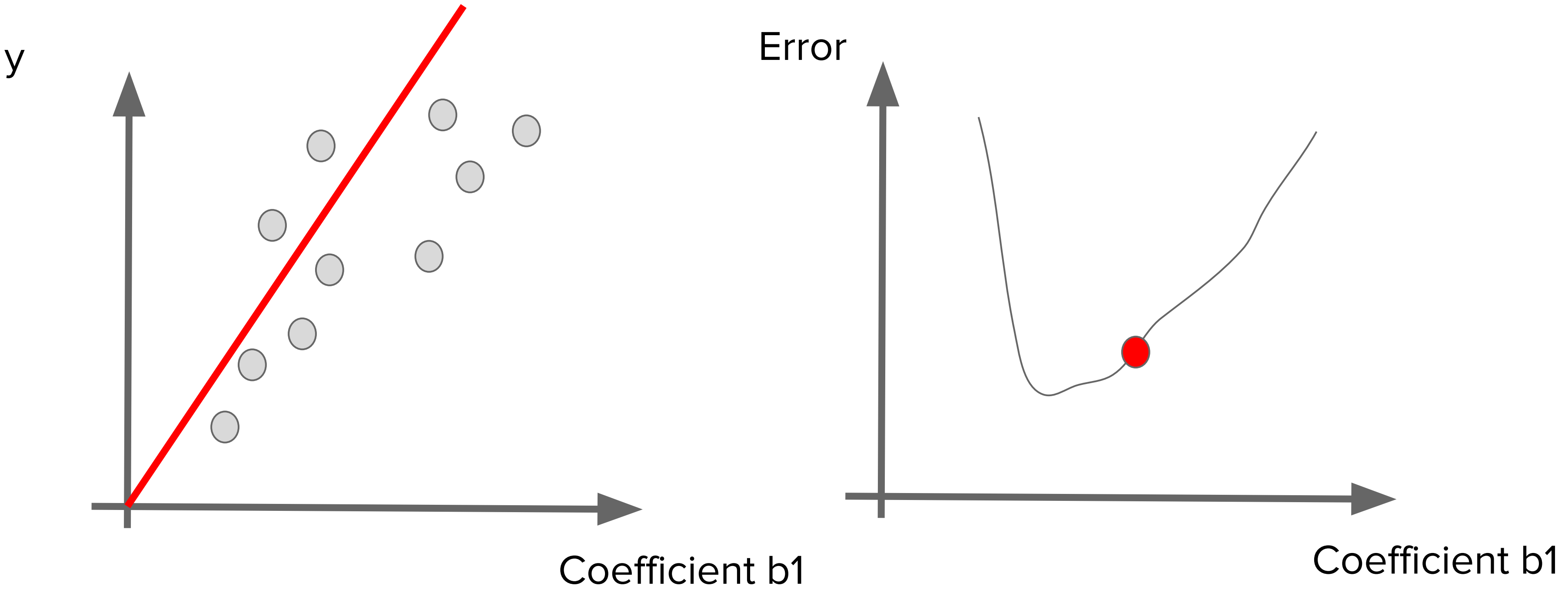

y
Coefficient b1
Error
Coefficient b1

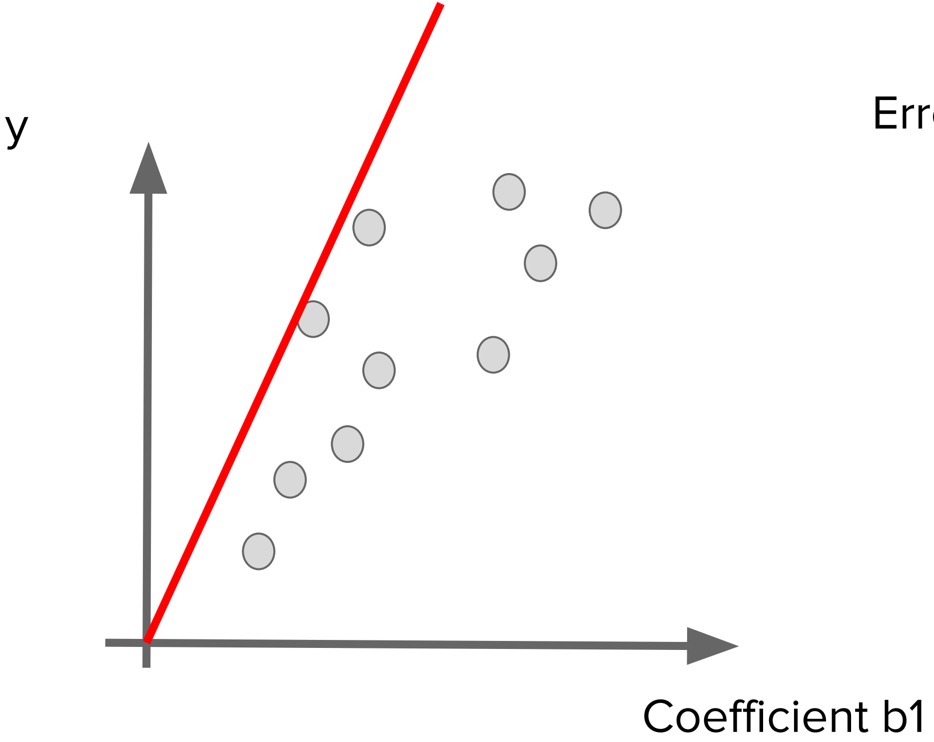

y
Coefficient b1

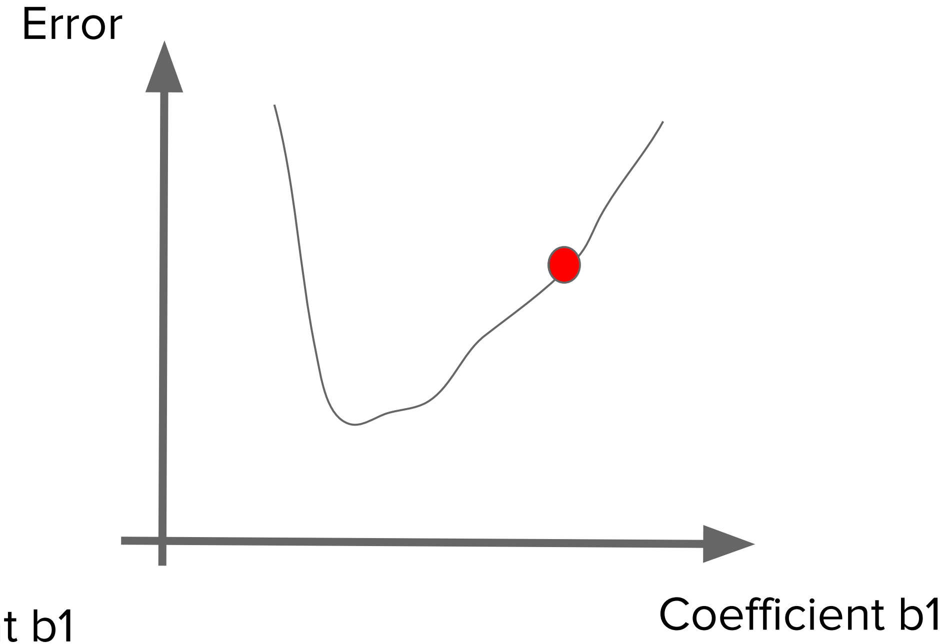

Error
Coefficient b1

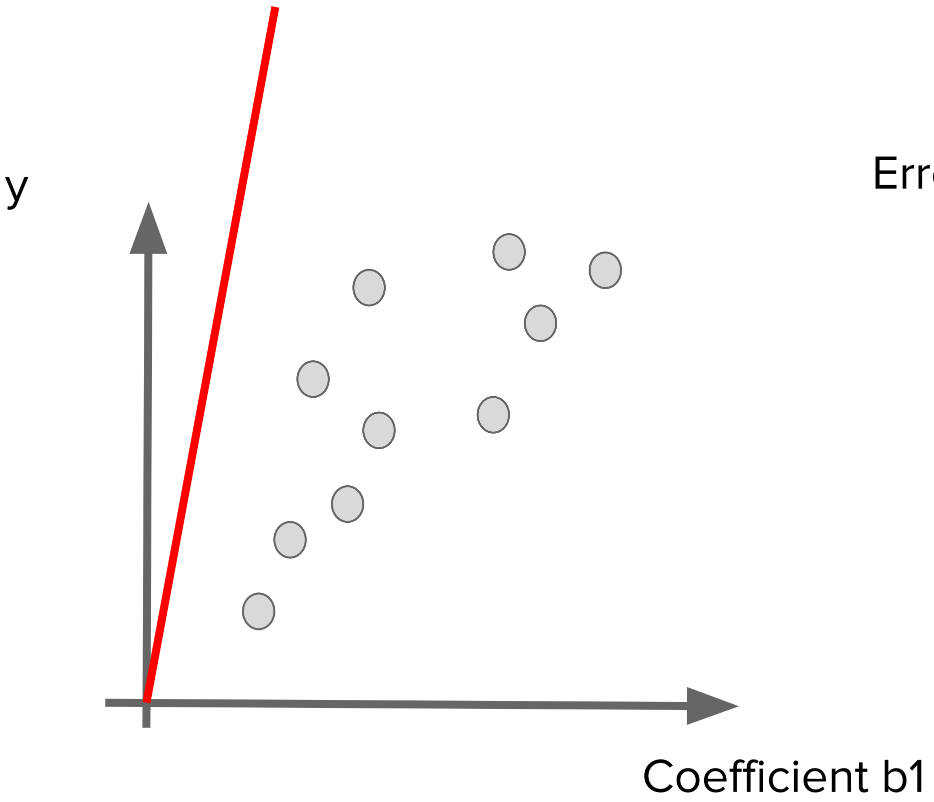

y
Coefficient b1

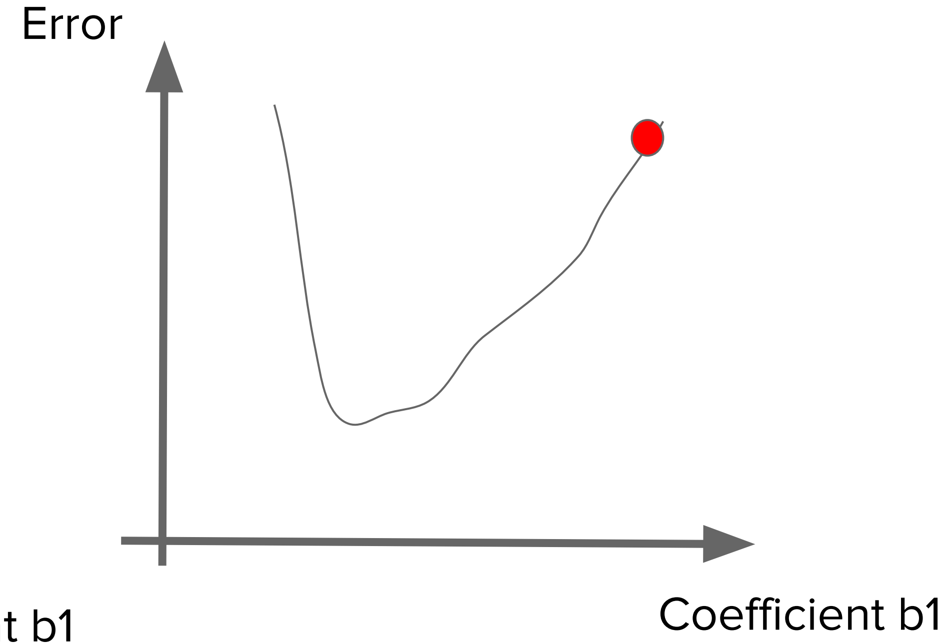

Error
Coefficient b1

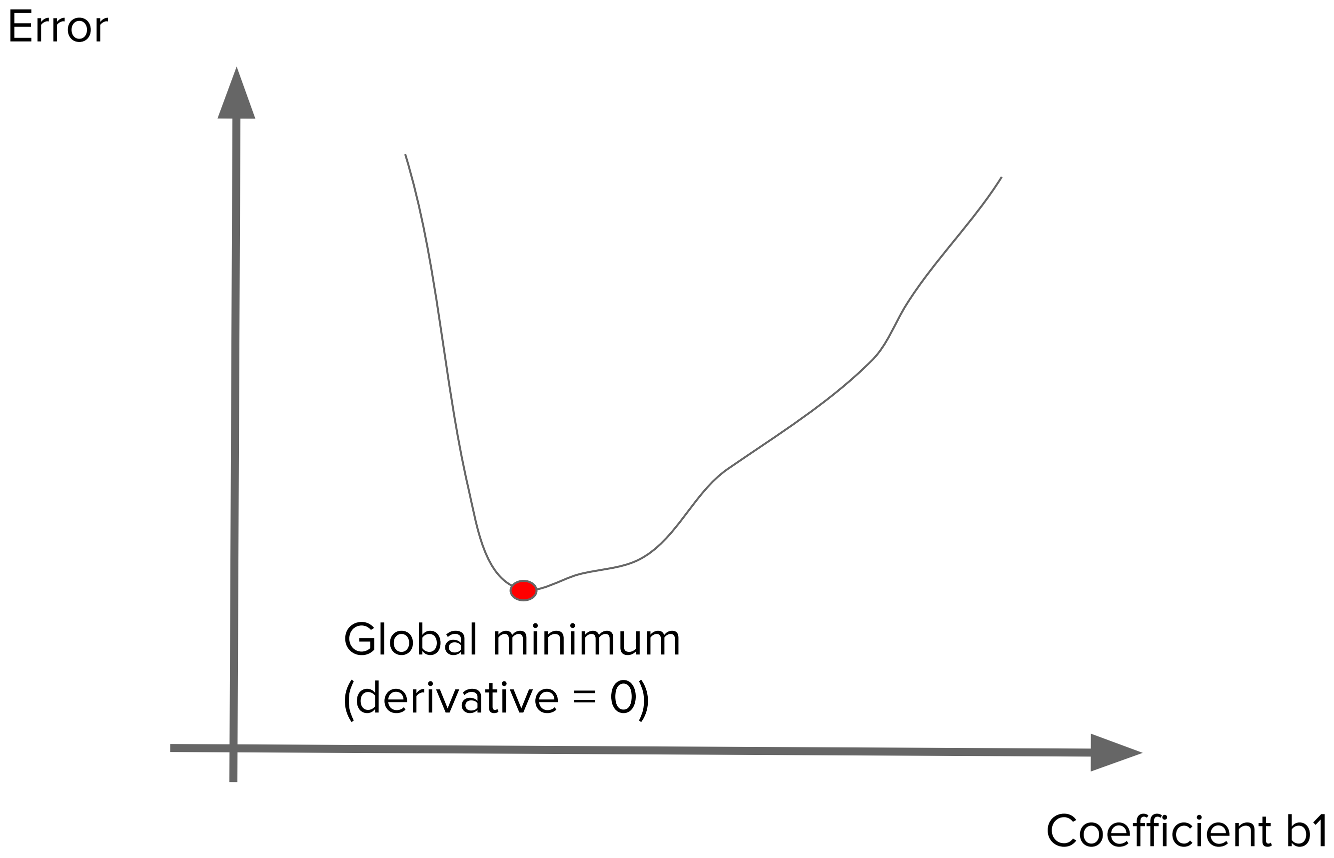

y
Coefficient b1
Error
Coefficient b1

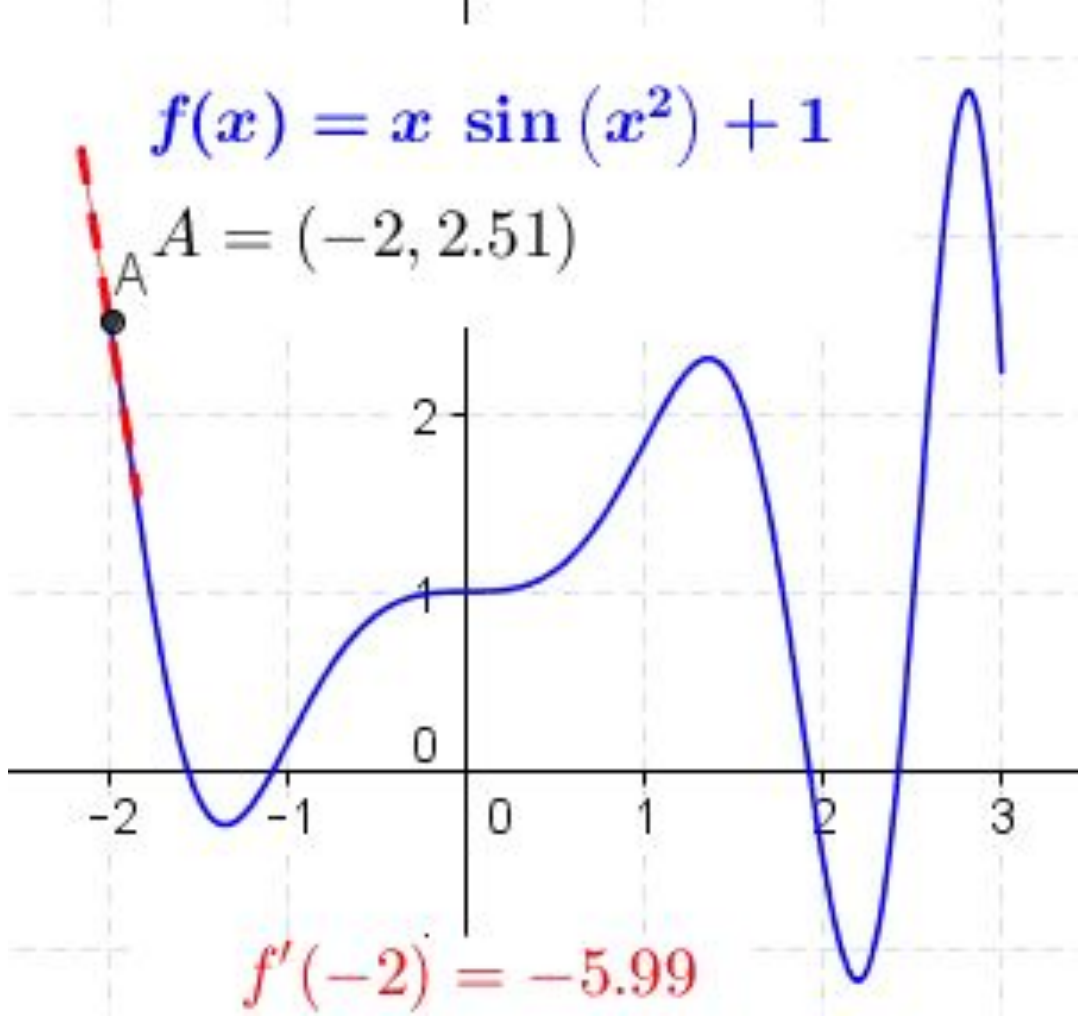

y
Coefficient b1


Error
Coefficient b1


y
Coefficient b1


Error
Coefficient b1


Error
Global minimum
(derivative = 0)
Coefficient b1


$f(x) = x \sin(x^2) + 1$
$A = (-2, 2.51)$
A
2
1
0
-2
-1
0
1
2
3
$f'(-2) = -5.99$

Error = SUM((y_hat - (b0+b1*x))**2)

Find derivative (it's just a function)

Set derivative = 0 ➔ find optimal (b0,b1)

Multiple Linear Regression

```
Salary =
+ b0
+ b1 * years_of_experience
+ b2 * ability_to_negotiate
- b3 * is_startup
+ b4 * is_finance
+ b5 * responsibilities
```

New assumptions:

independence of predictors

```
Salary =
+ b0
+ b1 * years_of_experience
+ b2 * months_of_experience
+ b3 * days_of_experience
```

Multicollinearity

(this is bad!)

Interpretation

What is this?

```
Salary =
+ b0
+ b1 * years_of_experience
+ b2 * ability_to_negotiate
- b3 * is_startup
+ b4 * is_finance
+ b5 * responsibilities
```

How many $ I get for each year of experience

Quick note on matrices

Matrices on python = Numpy arrays

Solve the exercise

or

Review basic linear algebra

(matrix multiplication, transpose, inverse, dot product, array-matrix multiplication, etc.)

Recap

1. R^2 - model vs. baseline
2. MLR - linear combination
3. Coefficients interpretation
4. Error = Loss = Function of b0,b1,b2,..,bn
5. Optimization = find the smallest error
6. !!! Review basic linear algebra !!!