

(Simple) Linear Regression

Harald (harald@ebi.ac.uk)

An Example

- Imagine you work in a pharmaceutical company which sells medical supplies to hospitals and doctors. You are interested in evaluating the effectiveness of a new advertisement program.

1. Start your virtual machine

2. Open the terminal and type

```
$ conda install matplotlib
```

3. Clone this repository

```
$ git clone https://github.com/sagar87/BTM\_2017.git
```

4. Change into the directory BTM_2017 and start jupyter notebook

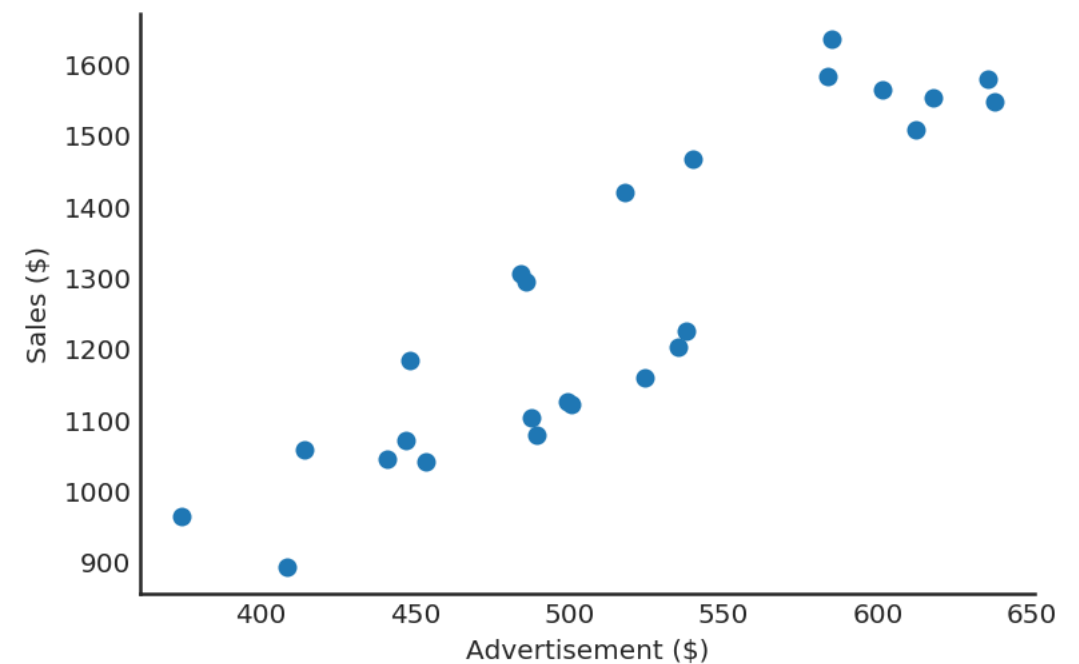
```
$ jupyter notebook
```

Problem 1

Load and plot the some data using numpy and matplotlib!

The model

- **Response Variable:** This is a random variable, it varies with changes in the predictor.
- **Predicting Variable:** This is a fixed variable, it does not change with the response and is fixed before the response is measured.



$$\underbrace{\hat{y}_i}_{\text{Response Variable}} = \hat{\beta}_0 + \hat{\beta}_1 \underbrace{x_i}_{\text{Predicting Variable}}$$

Why Regression ?

- **Prediction** of the response variable.
- **Modelling** the relationship between the response variable and the explanatory variable.
- **Testing** hypothesis of association relationships.

The Big Picture

- **Simple Linear Regression** $y = \beta_0 + \beta_1 x + \varepsilon,$
- **Logistic Regression (classification)** $F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$
- **Multivariate Linear Regression** $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \epsilon_i.$
- **Ridge Regression** $\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}$
- **Lasso** $\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$
- **Polynomial regression** $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$

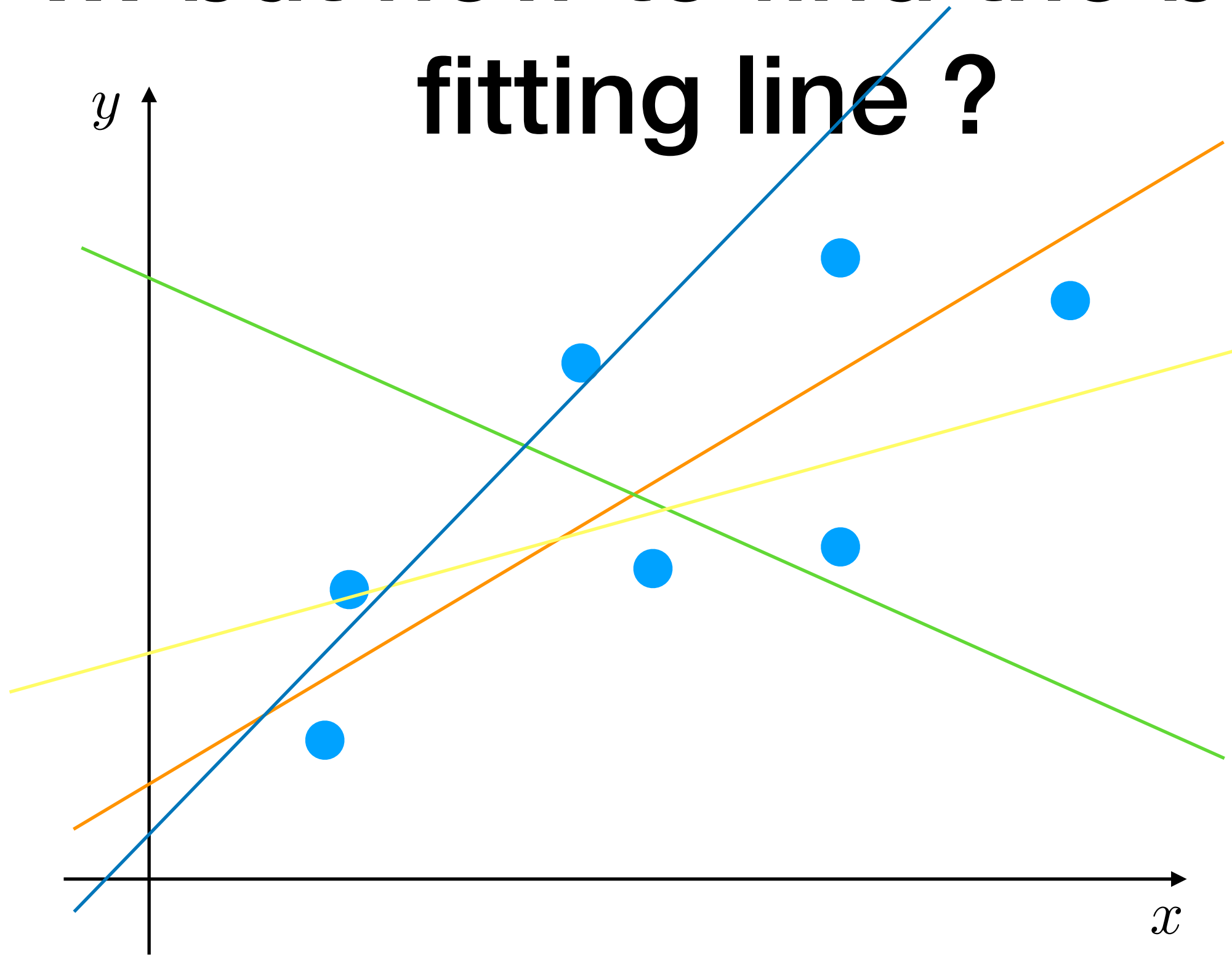
Simple Linear Regression

...

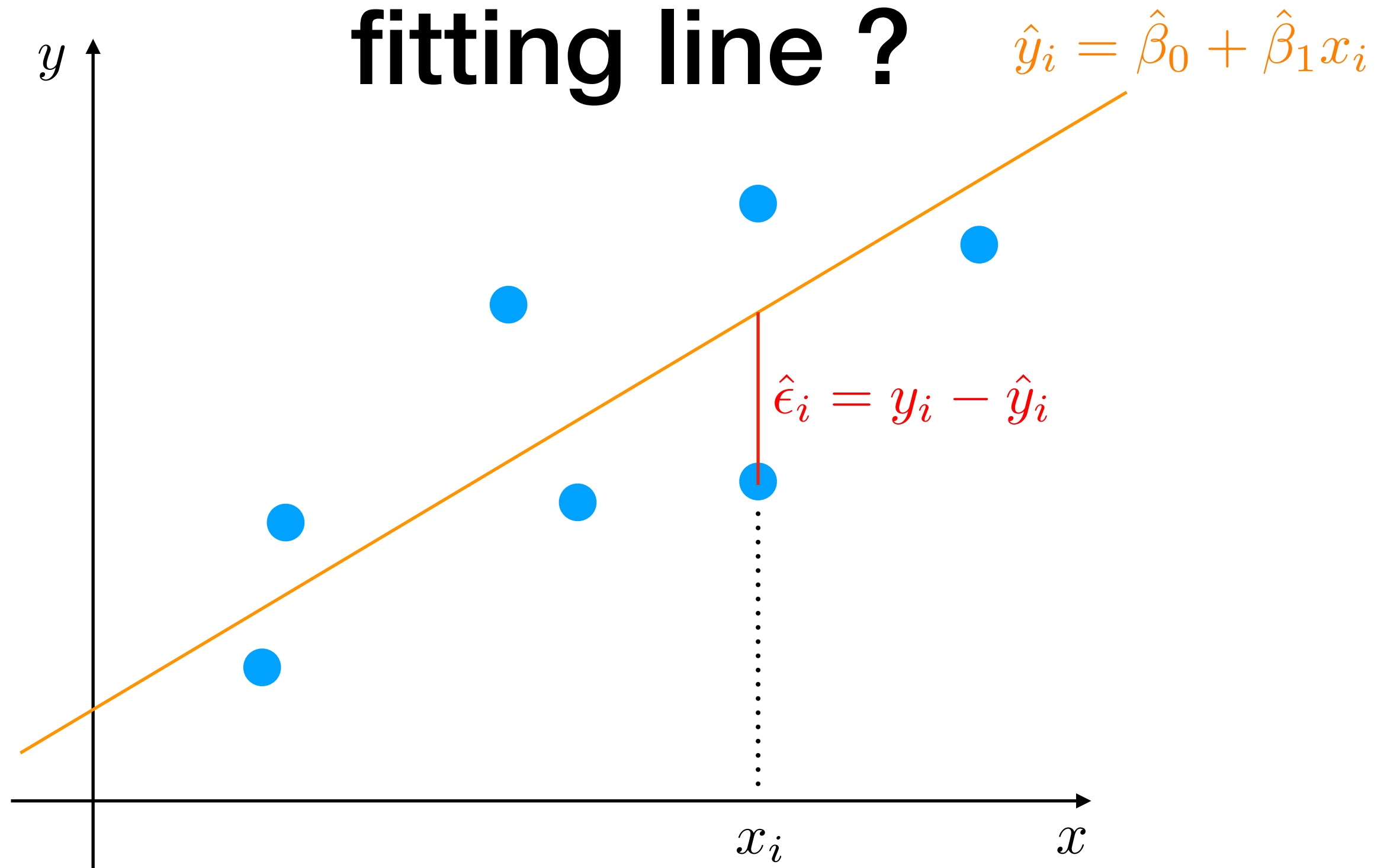
$$y_i = \underbrace{\hat{\beta}_0}_{\text{Intercept}} + \underbrace{\hat{\beta}_1}_{\text{Slope}} x_i + \underbrace{\epsilon_i}_{\text{Error}}$$

- Our goal is to find the best line describing the linear relationship between our variables of interest. In other words, we want to find the coefficients β_0 and β_1 .

... but how to find the best
fitting line ?

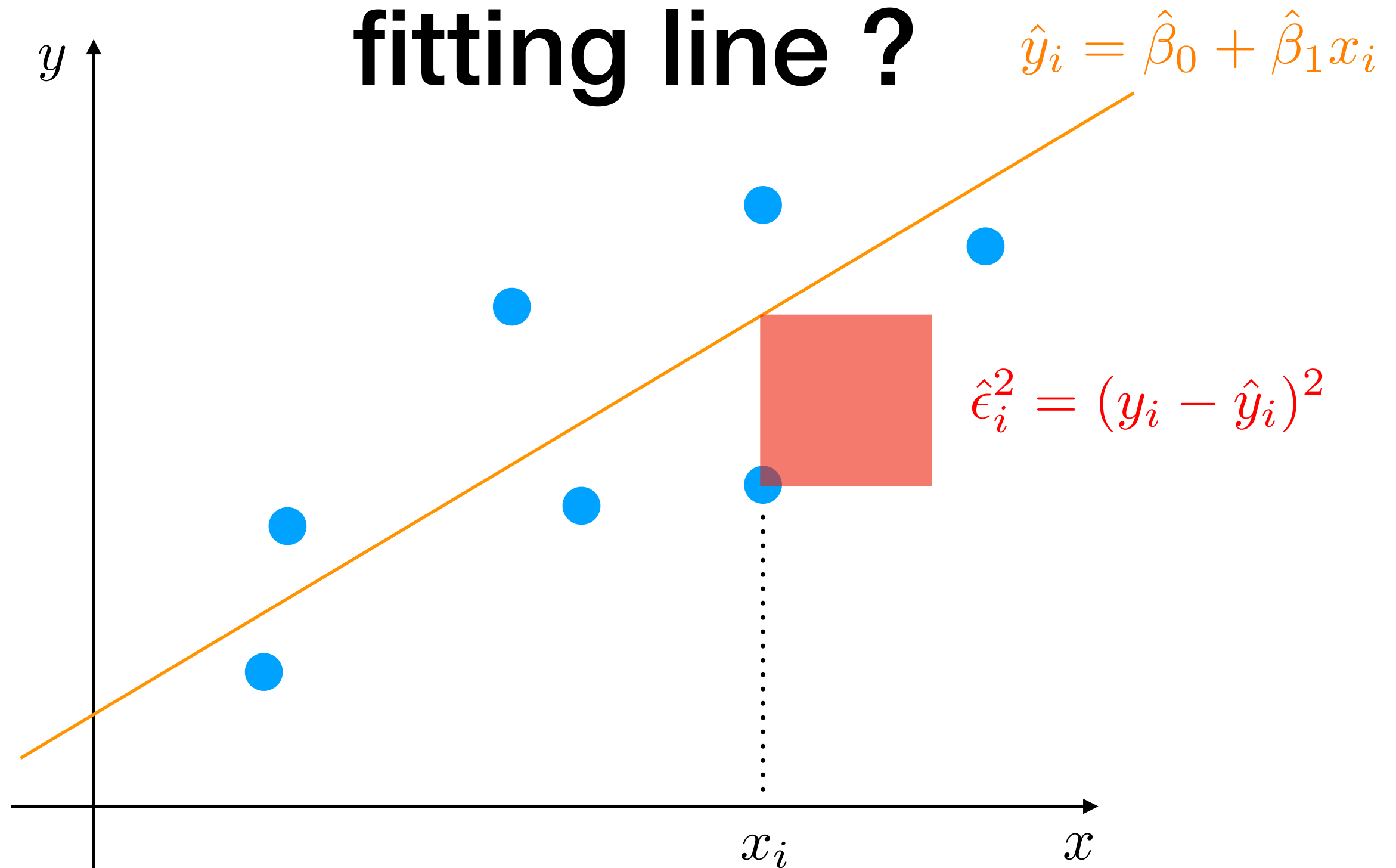


... but how to find the best
fitting line ?



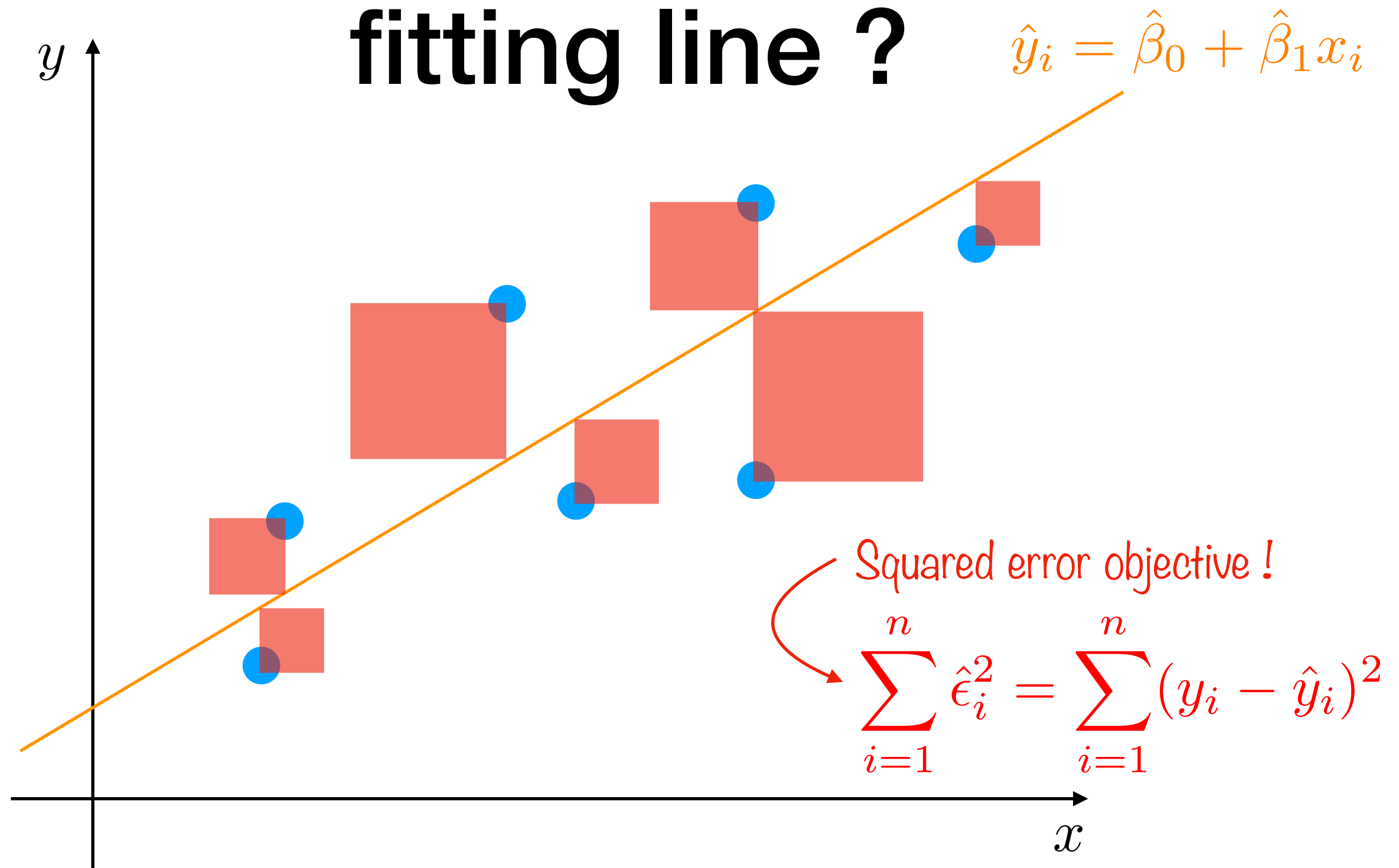
We can evaluate the quality of each possible regression line by choosing the one which gives us the smallest prediction error.

... but how to find the best
fitting line ?



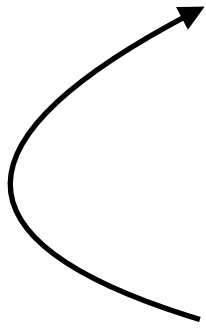
... actually it makes more sense to look at the squared
residuals rather than only at errors.

... but how to find the best fitting line ?



In order to come up with the optimal regression coefficients we need to consider the total error.

... but how to find the best fitting line ?

$$\operatorname{argmin}_{(\hat{\beta}_0, \hat{\beta}_1)} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x))^2$$


From calculus we remember that we can solve such a problem by finding the first derivative of the function and setting it to zero.

... but how to find the best fitting line ?

$$\operatorname{argmin}_{(\hat{\beta}_0, \hat{\beta}_1)} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = 0$$

$$-2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i = 0$$

$$n\hat{\beta}_0 = \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\beta}_1 x_i$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_i$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

.....

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = 0$$

$$2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$$

$$-\sum_{i=1}^n y_i x_i + \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$

$$-\sum_{i=1}^n y_i x_i + (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$

$$-\sum_{i=1}^n y_i x_i + \bar{y} \sum_{i=1}^n x_i - \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$

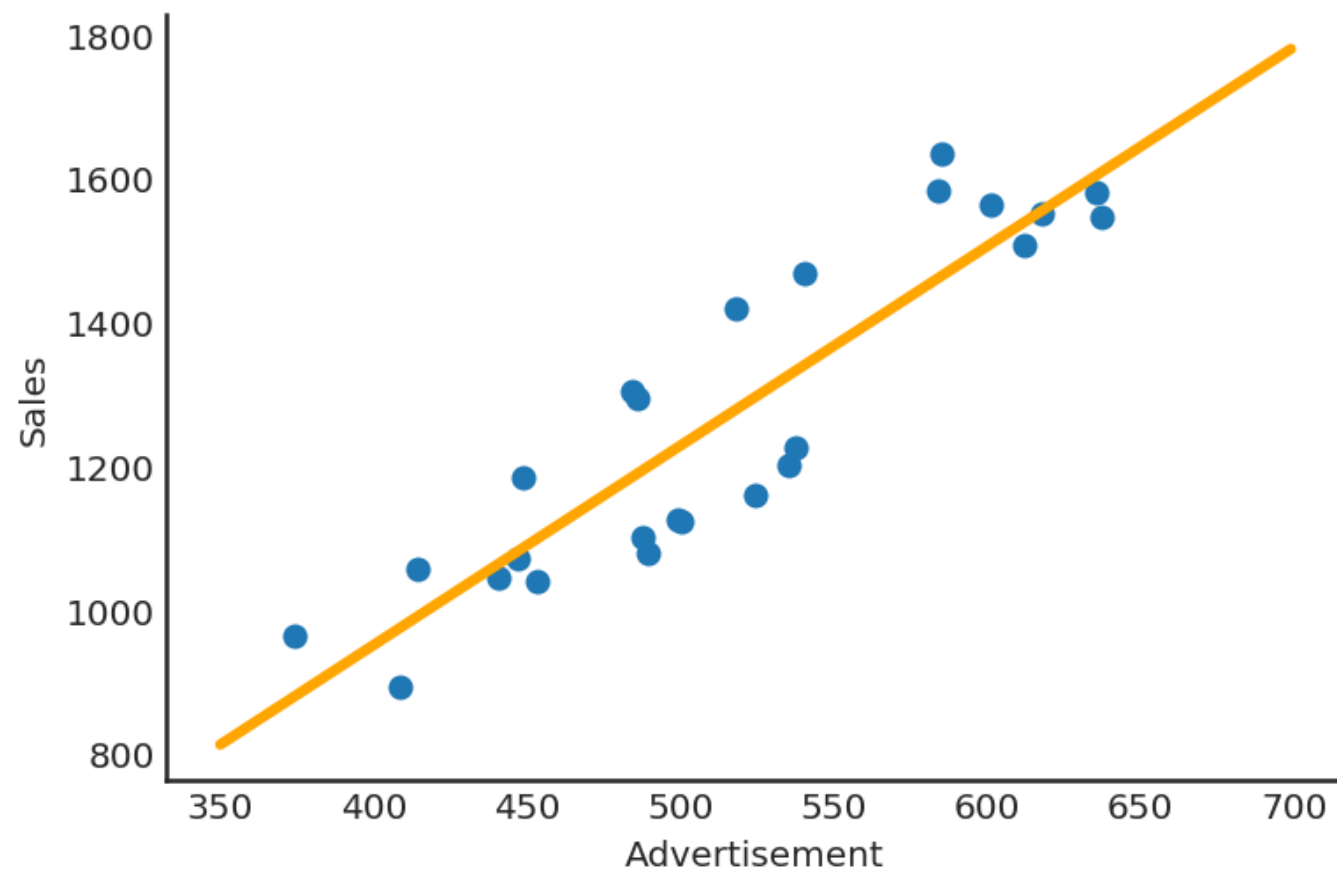
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \bar{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i}$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Problem 2

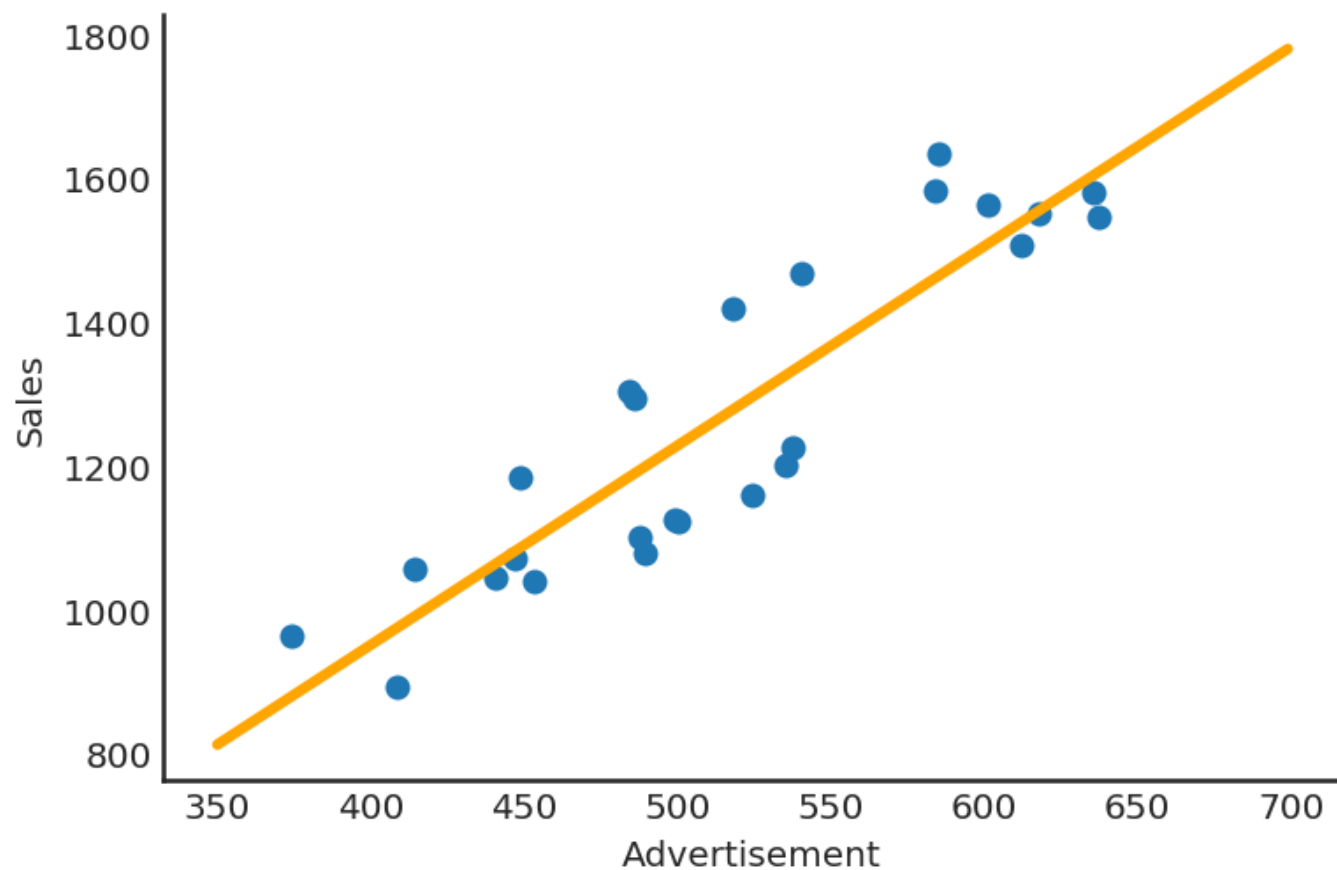
Fit a linear regression model. What are the estimated coefficients?

Interpretation of model parameters



- β_0 : Estimated expected value of the response when the predicting variable is 0
- β_1 : Estimated expected change of the response variable when the predicting variable increases by one unit

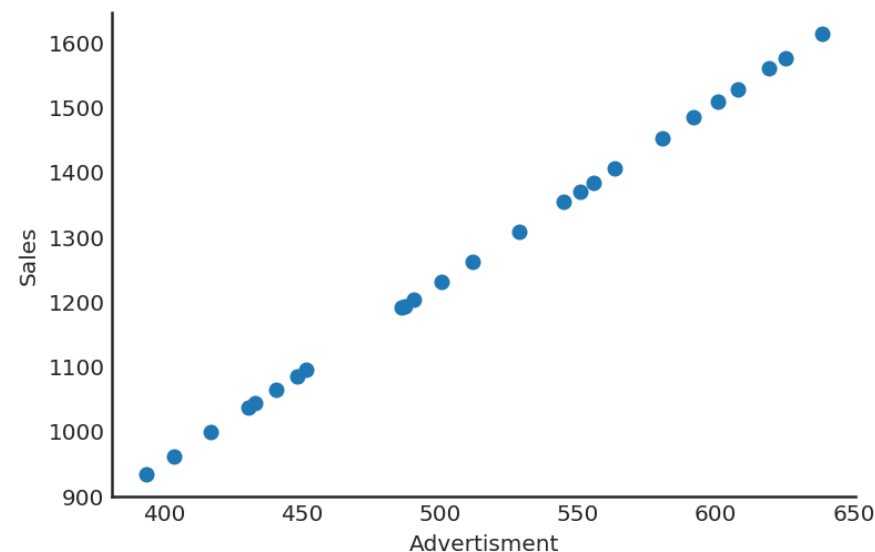
Interpretation of model parameters



- $\beta_1 \geq 0$ positive relationship
- $\beta_1 \leq 0$ negative relationship
- $\beta_1 = 0$ no relationship

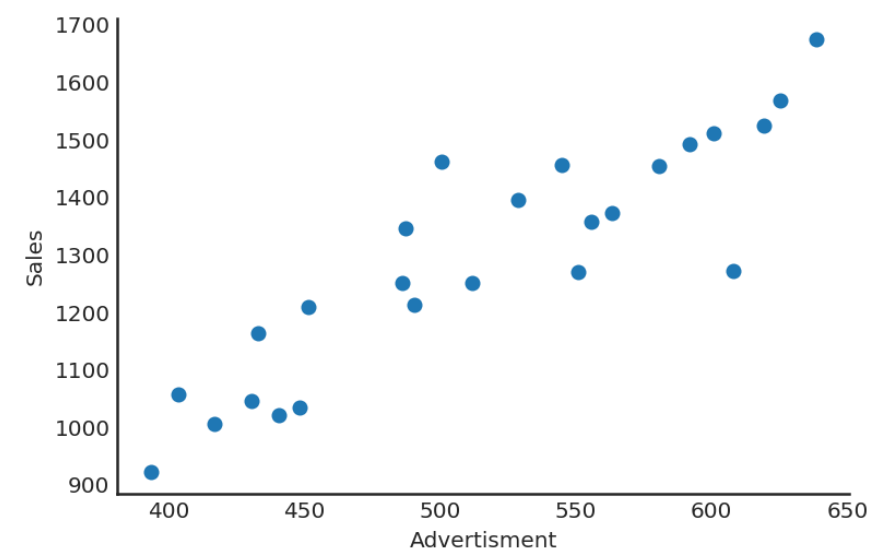
Functional vs Statistical Relationships

$$y_i = \beta_0 + \beta_1 x_i$$



```
from scipy.stats import uniform, norm
f = lambda x: model.intercept_[0] + model.coef_[0]*x
x = uniform.rvs(loc = 374, scale= 300, size=25)
y = f(x)
...
```

$$Y_i \sim \mathcal{N}(\underbrace{\beta_0 + \beta_1 x_i}_{\text{True relationship}}, \sigma^2)$$

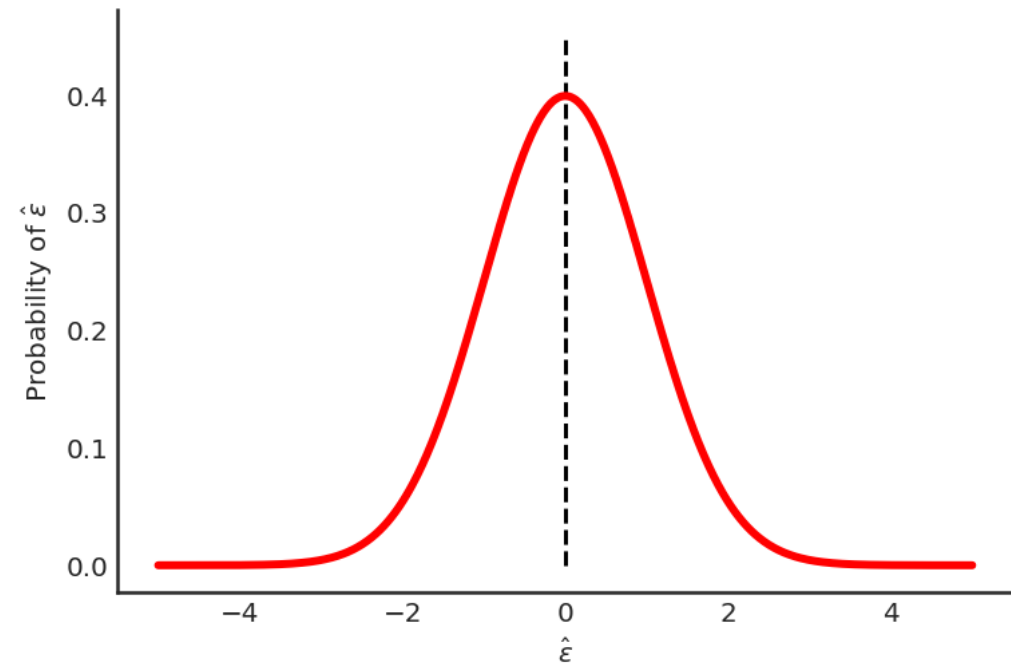
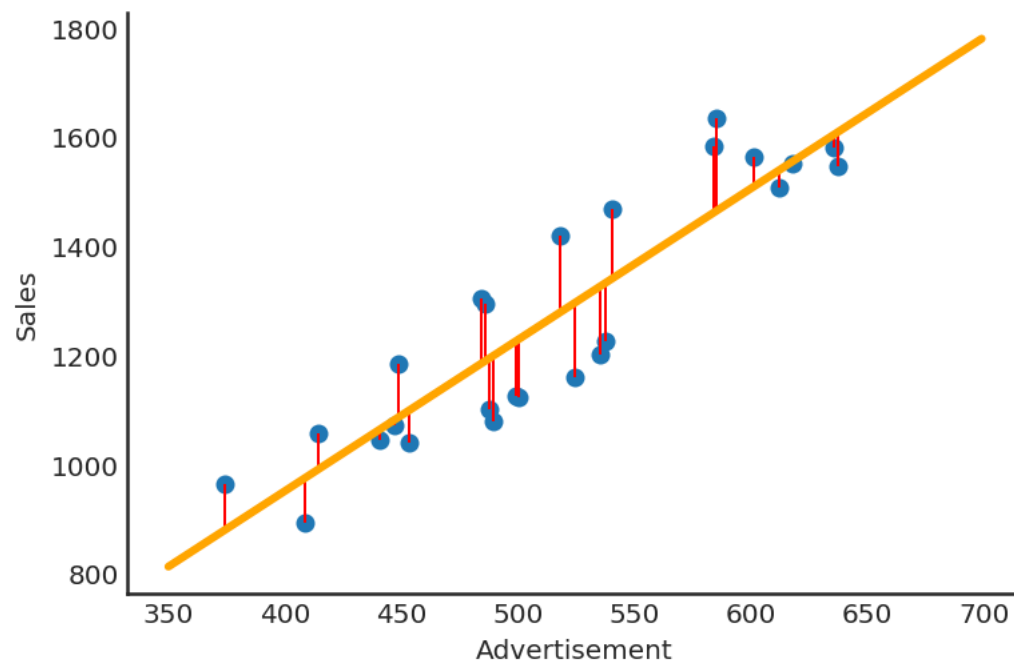


```
from scipy.stats import uniform, norm
f = lambda x: model.intercept_[0] + model.coef_[0]*x
x = uniform.rvs(loc = 374, scale= 300, size=25)
y = norm.rvs(loc = f(x), scale=101)
...
```

By performing linear regression we try to estimate the **mean** of Y_i which is a **linear** function of x_i !

$$\mathbb{E}(Y_i) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Probabilistic Assumptions



- Zero Mean Assumption
- Constant Variance Assumption
- IID Assumption
- Normal Assumption

$$\mathbb{E}(\epsilon) = 0$$

$$\mathbb{V}(\epsilon) = \sigma^2$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Let's make the picture complete ...

- We have seen

$$Y_i \sim \mathcal{N}\left(\underbrace{\beta_0 + \beta_1 x_i}_{\text{True relationship}}, \sigma^2\right) \quad \text{and} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

we can estimate the mean of Y_i

$$\mathbb{E}(Y_i) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

but can we also find an estimate $\hat{\sigma}^2$ for σ^2 ?

Let's make the picture complete ...

- We have seen

$$Y_i \sim \mathcal{N}\left(\underbrace{\beta_0 + \beta_1 x_i}_{\text{True relationship}}, \sigma^2\right) \quad \text{and} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

we can estimate the mean of Y_i

$$\mathbb{E}(Y_i) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

but can we also find an estimate $\hat{\sigma}^2$ for σ^2 ?

It turns out we can :) $\hat{\sigma}^2 = \text{MSE} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$

Problem 3

Think hard about the MSE!

Inference on β_1

- One can show: $\mathbb{E}(\hat{\beta}_1) = \beta_1$
- Also realise that
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} y_i = \sum_{i=1}^n \underbrace{\frac{(x_i - \bar{x})}{S_{xx}}}_{c_i} y_i = \sum_{i=1}^n c_i y_i$$

implying that $\hat{\beta}_1$ is a linear combination of Y. If we also assume that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ then

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

However, we don't know σ^2 but we can replace it with the MSE, then the sampling distribution becomes a t-dist. with $n-2$ df

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{MSE}}{S_{xx}}}} \sim t_{n-2}$$

Challenge! An alpha-level hypothesis test for β_1

1. We specify

1. Null hypothesis: $H_0 : \beta_1 = 0$

2. Alternative hypothesis: $H_A : \beta_1 \neq 0$

2. We calculate the test statistic $t^* = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{\sum_i^n (x_i - \bar{x})^2}}}$

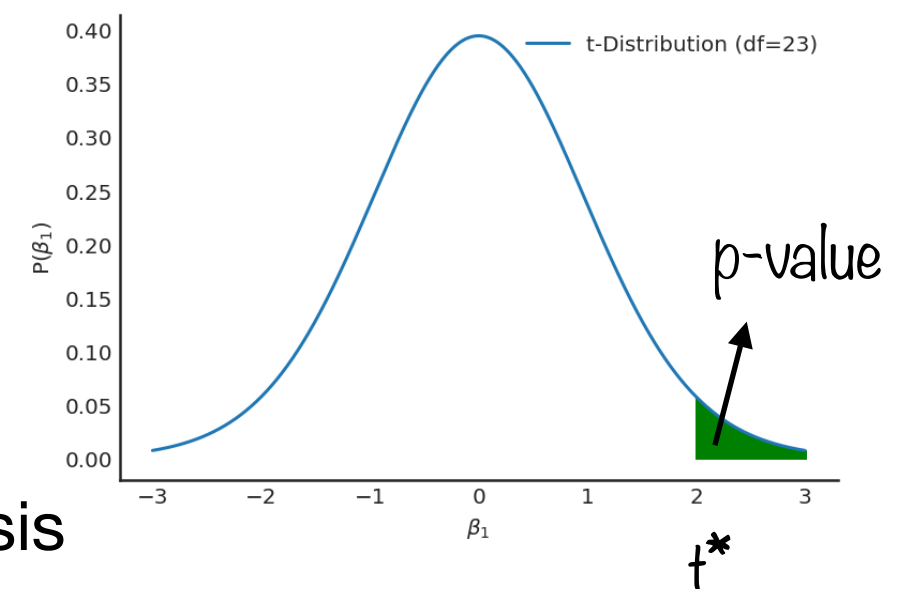
3. We use t^* to compute the P-value

4. We decide

1. if P-value < alpha we reject the null hypothesis

2. if P-value > alpha we fail to reject the null hypothesis

How likely is it that we'd get a statistic t^* as extreme as we did if the null hypothesis were true.



Hint: This distribution is symmetric.

Problem 4

Calculate the t-statistic for β_1 and report its p-value !
Should we reject the null hypothesis ($\alpha=0.01$)?