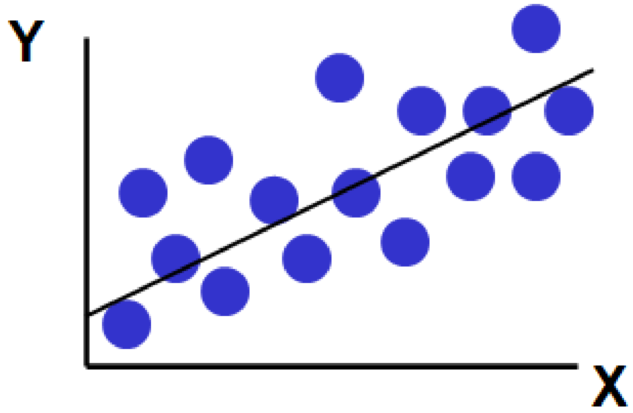
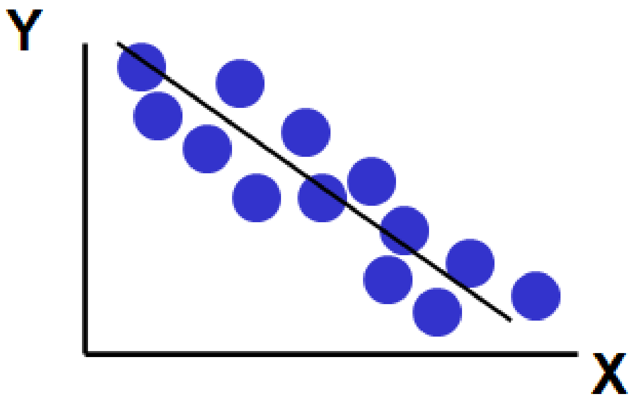
General Linear Model

Part 3.3 Linear Regression, Correlation, Hypothesis Testing

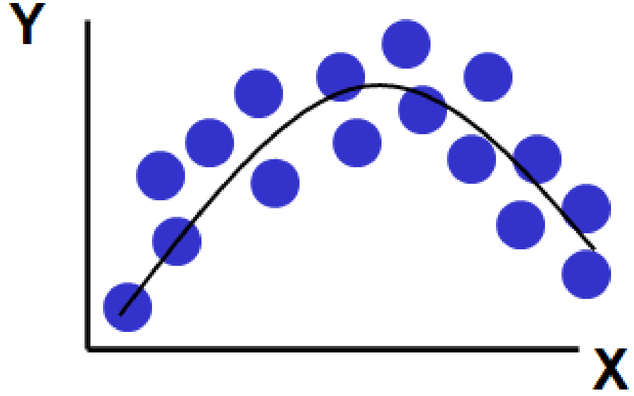
Types of Relationships

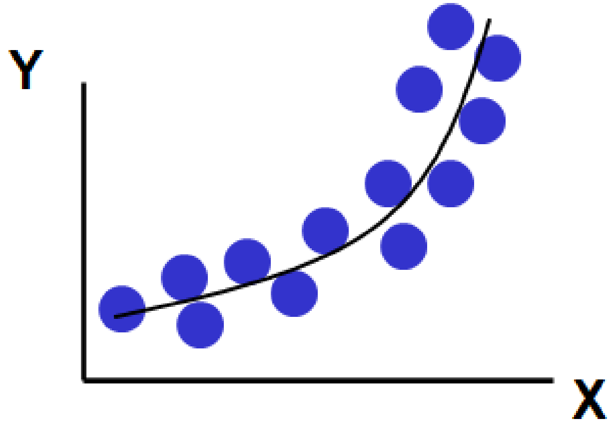
Linear relationships





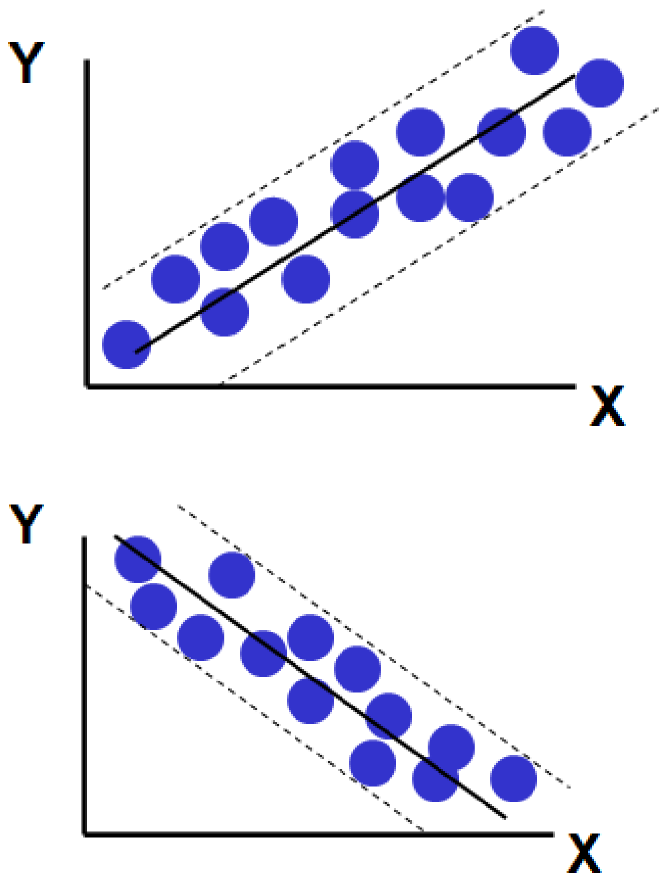
Nonlinear relationships



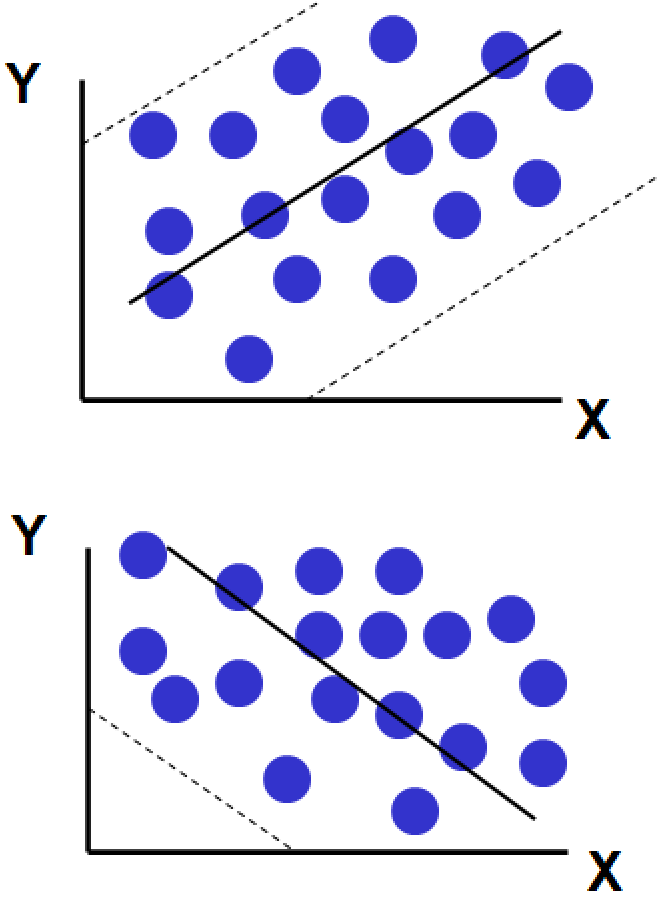


Types of Relationships

Strong relationships

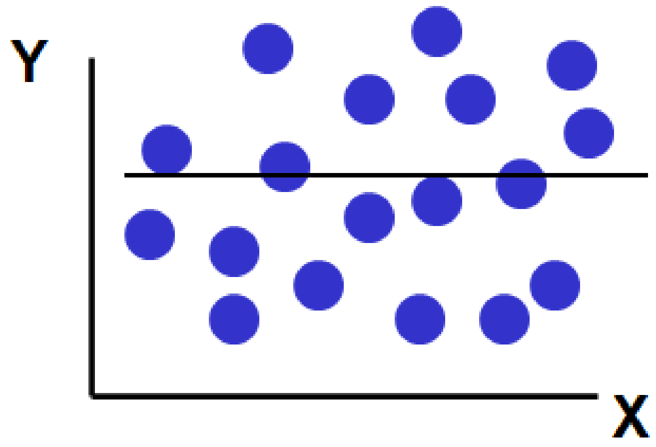


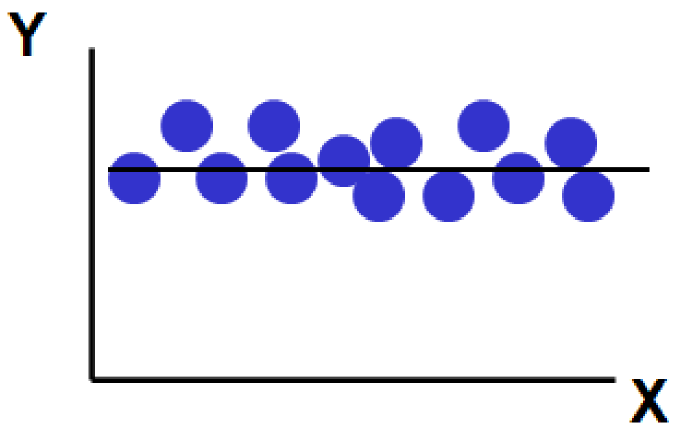
Weak relationships



Types of Relationships

No relationship





Relationships Between Variables

Some examples:

1. X = Advertisement, Y = Sales
2. X = Money growth, Y = Inflation
3. X = Income, Y = Willingness to pay for a car

Goal:

Explain relationship between X and Y

Focus:

Linear relationships

Linear Regression: Roadmap

Simple (Linear) Regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

(ε : unknown error term)

Multiple (Linear) Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

(ε : unknown error term)

The approaches we have used so far (normal distribution, point estimation, interval estimation, hypothesis testing, … …) are applied to each β_j and to ε.

Simple Linear Regression

One outcome variable: Y

One predictor variable: X

Data (paired): (X, Y)

The basic ideas

- Y as a linear function of X
- Explain the impact of changes in X on Y
- Predict the value of Y based on a given value of X

Example: Car Dealership

A key piece of information for securing the sale of a car is understanding the willingness to pay of the potential customer.

This helps the salesperson focusing on a certain range of car makes and models that might secure the sale at the best possible competitive price.

How can income information be used?

→ Use it to predict customers' spending?

→ How accurate is the spending prediction?

Example: Car Dealership

1. What are we modeling?

2. How do we interpret the model?

Population Model

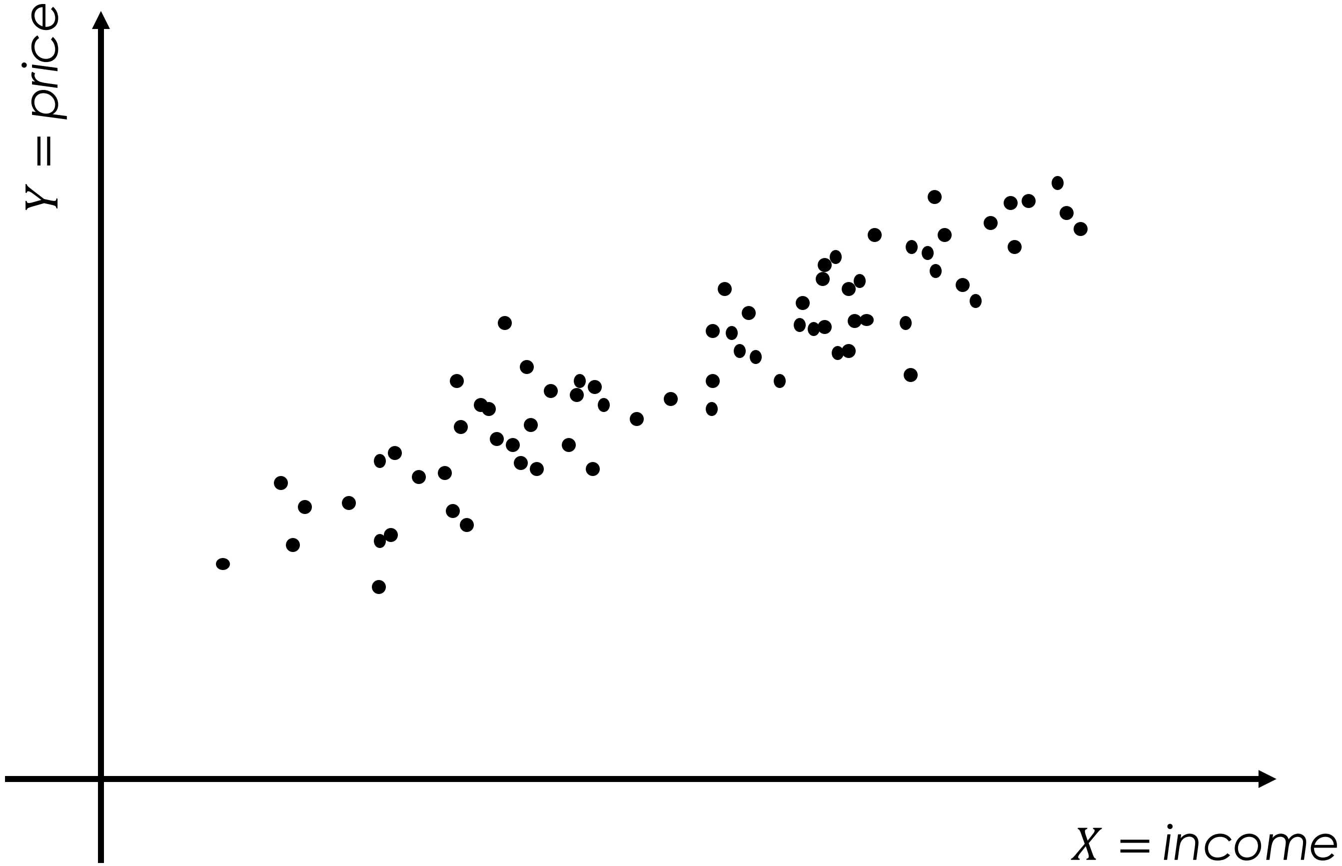
… describes our assumptions about the population relationship between Y and X.

Assumptions:

1. Y and X are linearly related (sometimes called the tendency).
2. Individual observations can differ from the tendency in a random way (These differences are called error terms).
3. The error terms follow a standard normal distribution.

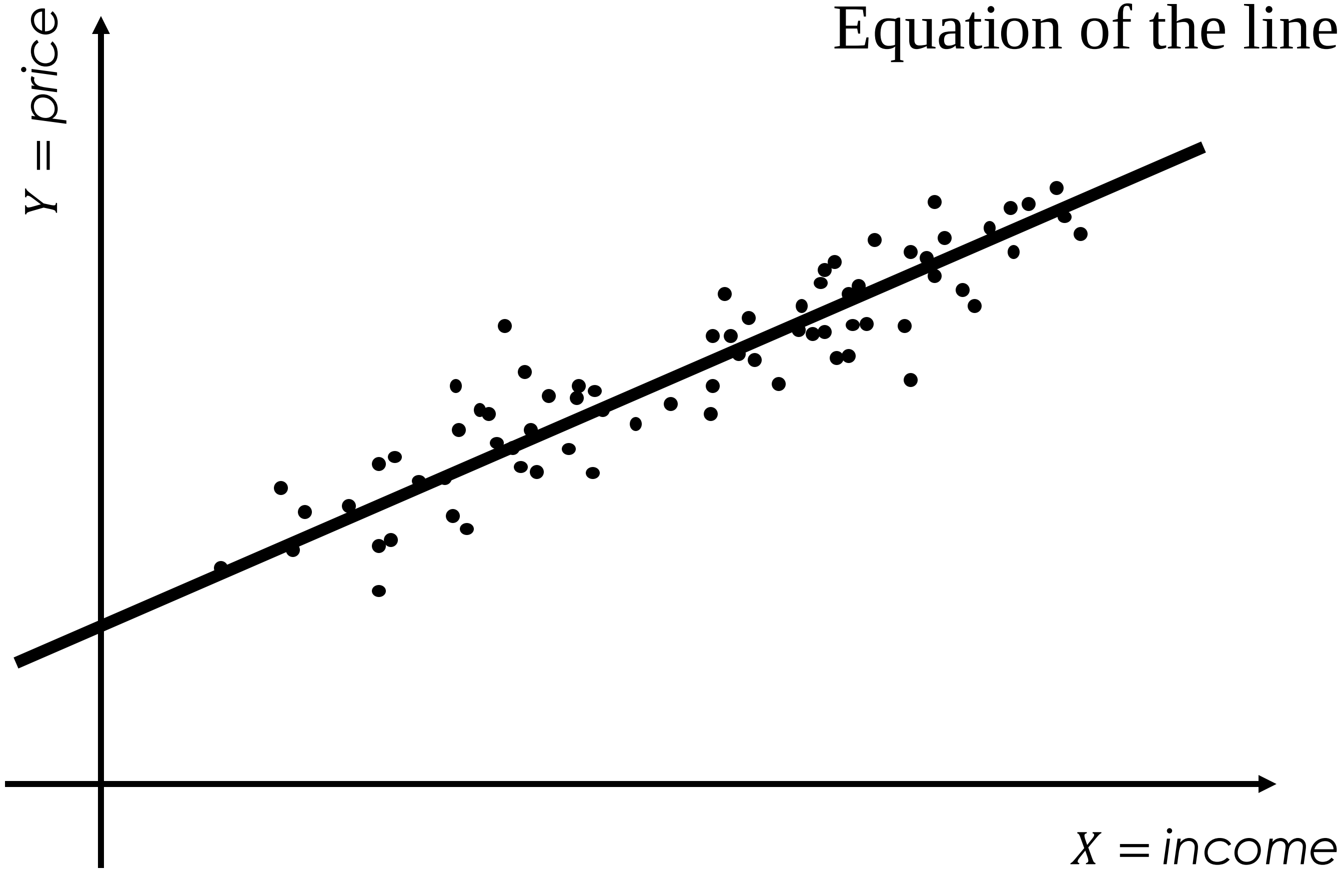
Let's look at some pictures to show these assumptions!

Simple Linear Regression: Population



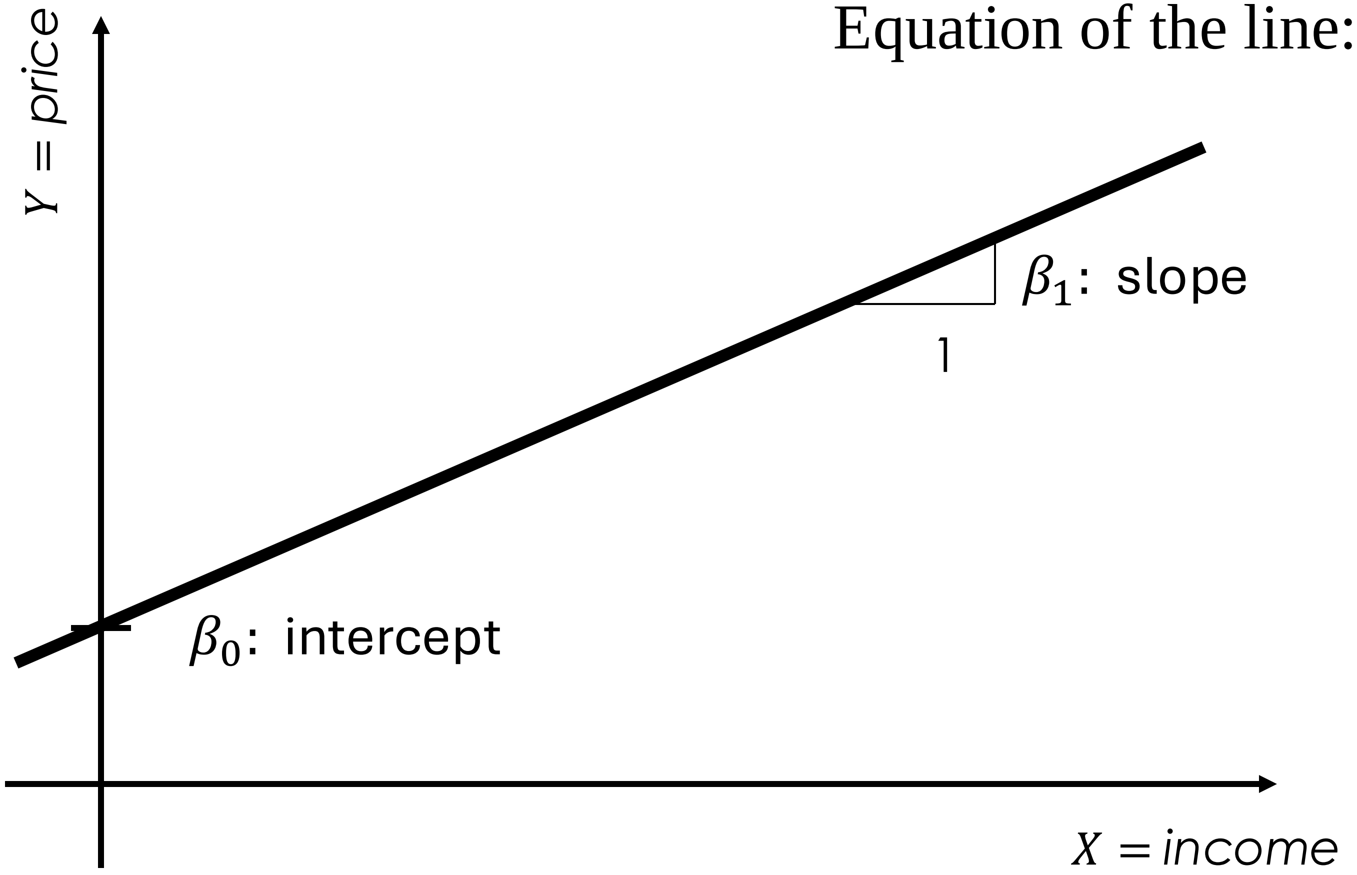
Simple Linear Regression: Tendency

Equation of the line: $\beta_0 + \beta_1 X$

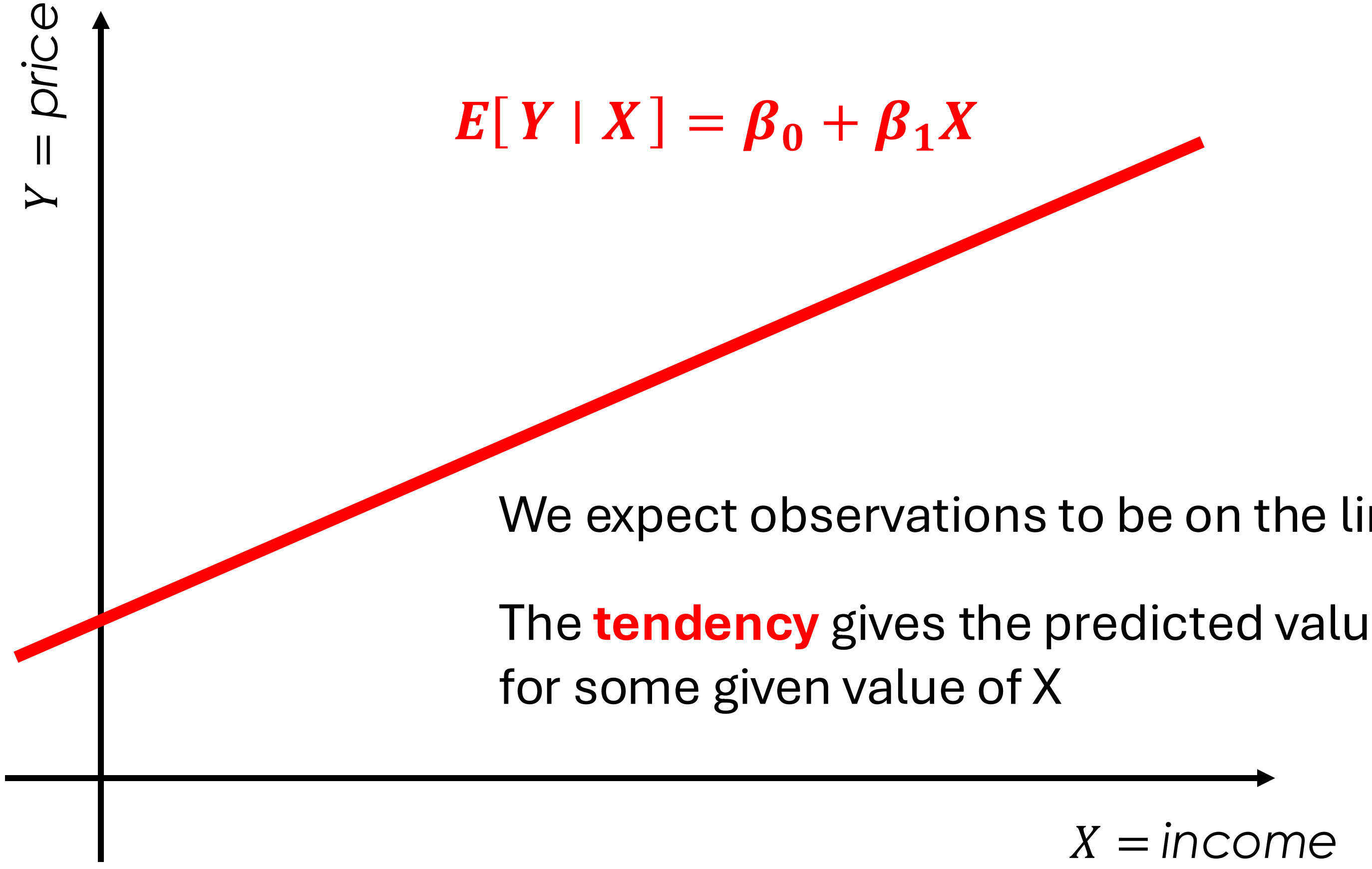


Simple Linear Regression: Tendency

Equation of the line: $\beta_0 + \beta_1 X$



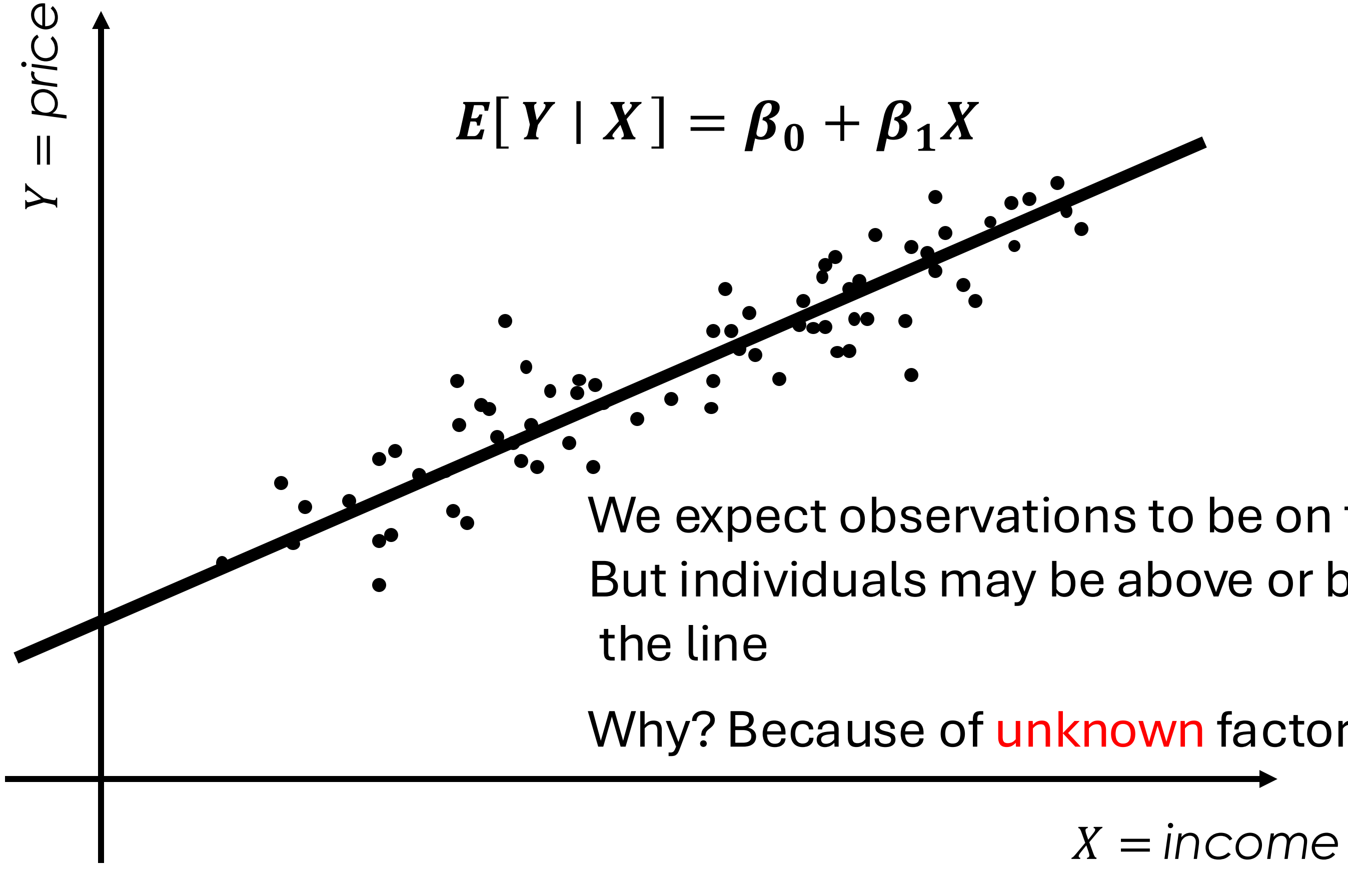
Simple Linear Regression: Tendency



We expect observations to be on the line.

The **tendency** gives the predicted value of Y for some given value of X

Simple Linear Regression

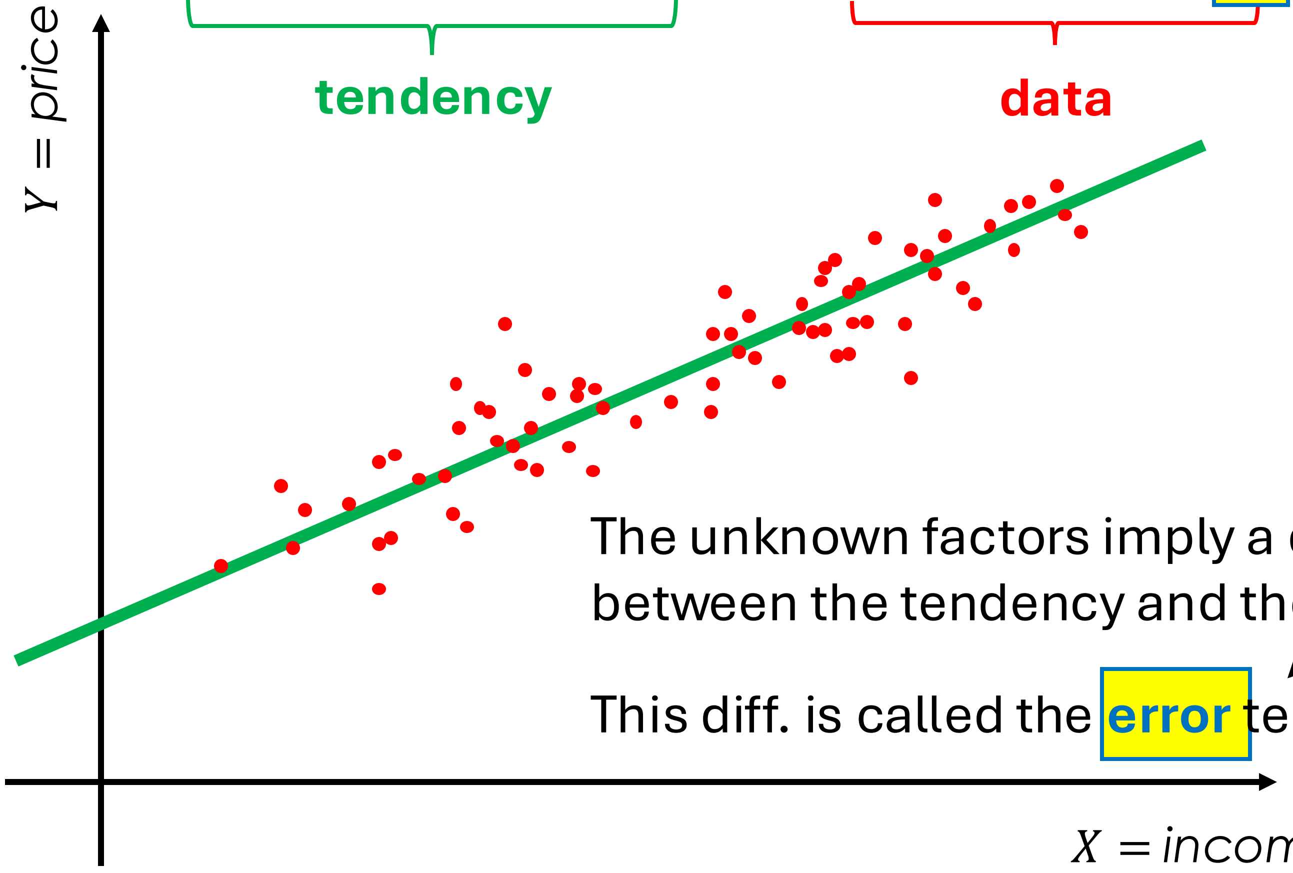


We expect observations to be on the line.
But individuals may be above or below the line

Why? Because of unknown factors

Simple Linear Regression: The Error

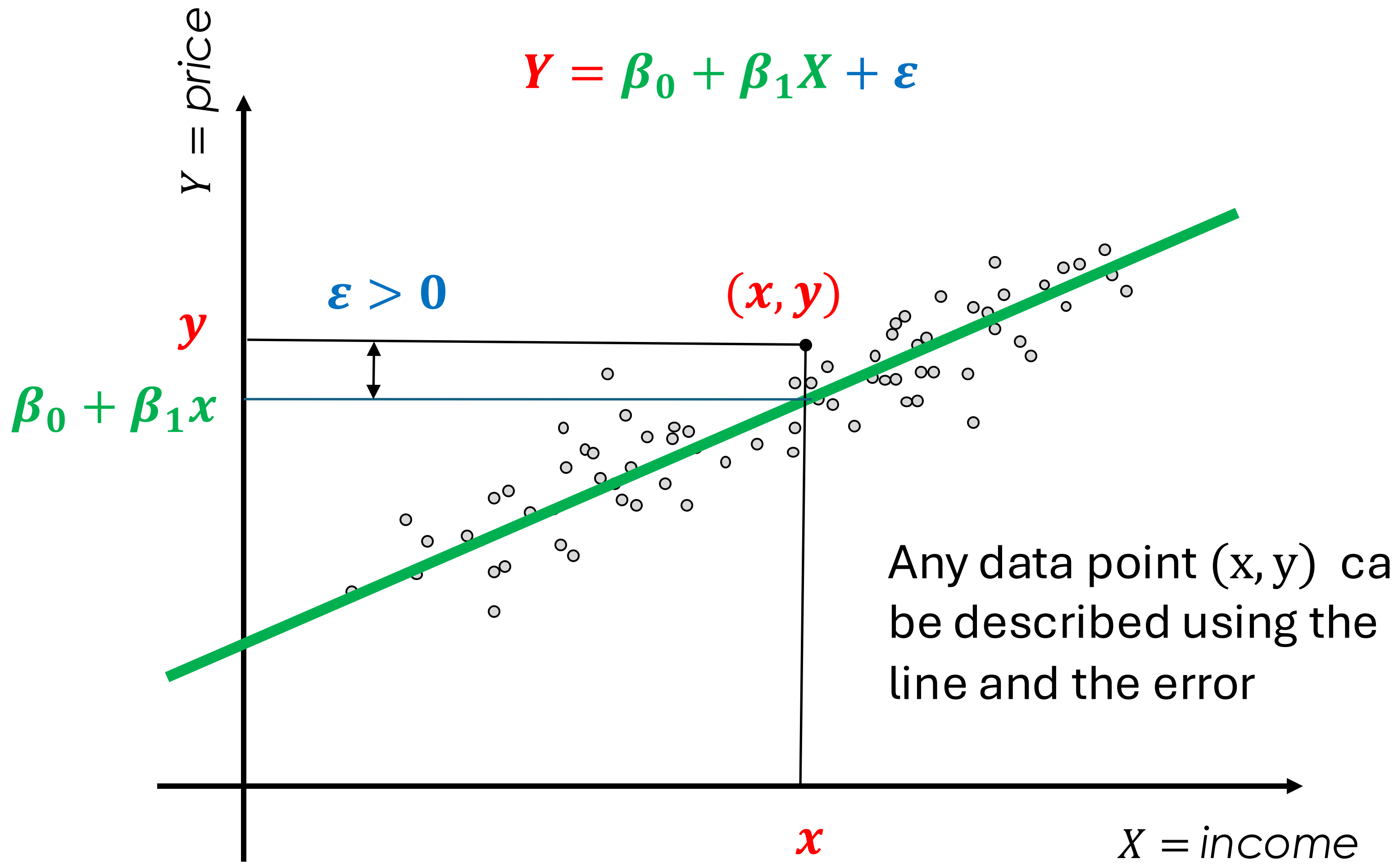
$$\underbrace{E[\,Y \mid X\,] = \beta_0 + \beta_1 X}_{\text{tendency}} \qquad \underbrace{Y = \beta_0 + \beta_1 X + \boxed{\boldsymbol{\varepsilon}}}_{\text{data}}$$



The unknown factors imply a difference between the tendency and the data.

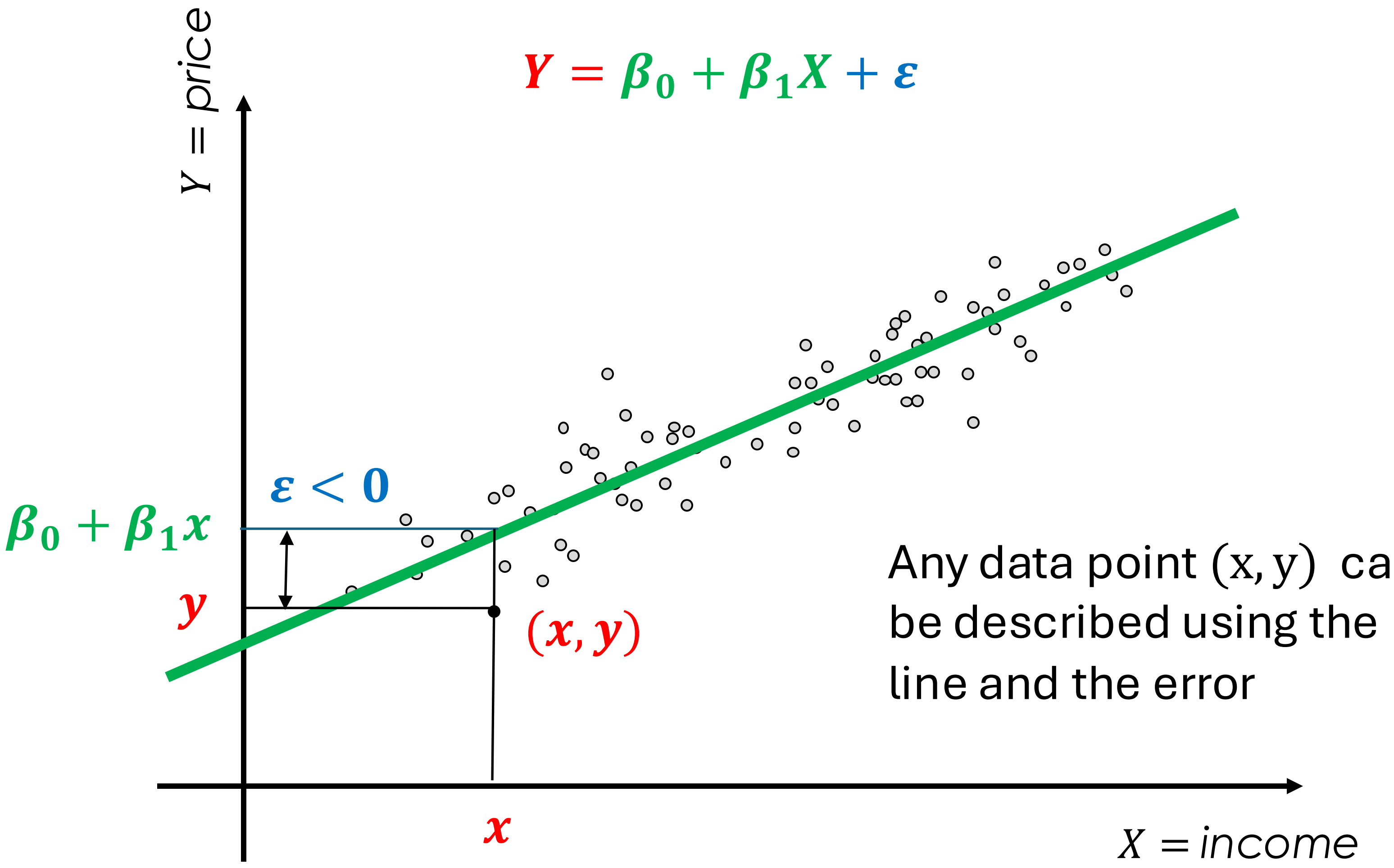
This diff. is called the **error** term

Simple Linear Regression: The Error



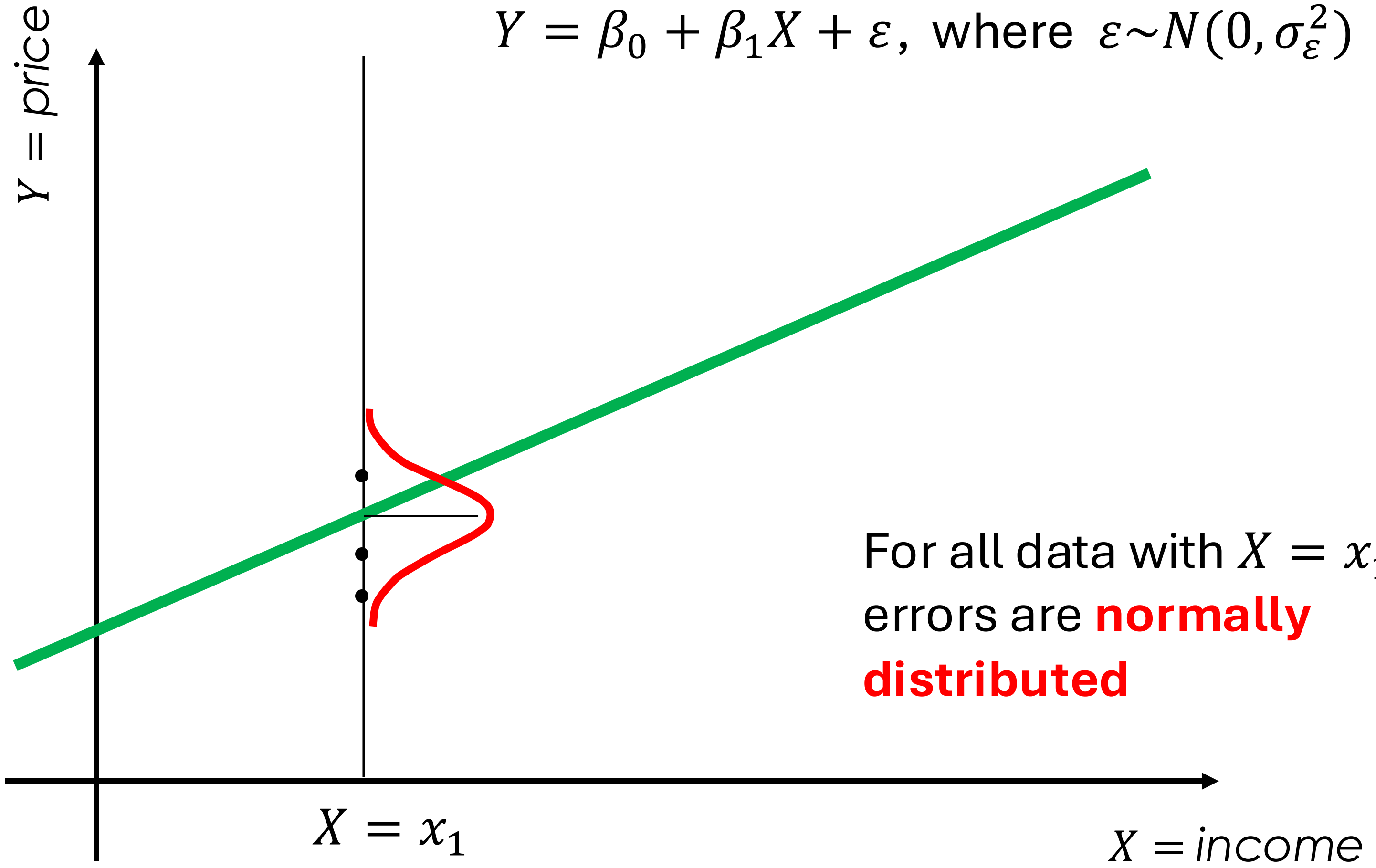
Any data point (x, y) can be described using the line and the error

Simple Linear Regression: The Error



Any data point (x, y) can be described using the line and the error

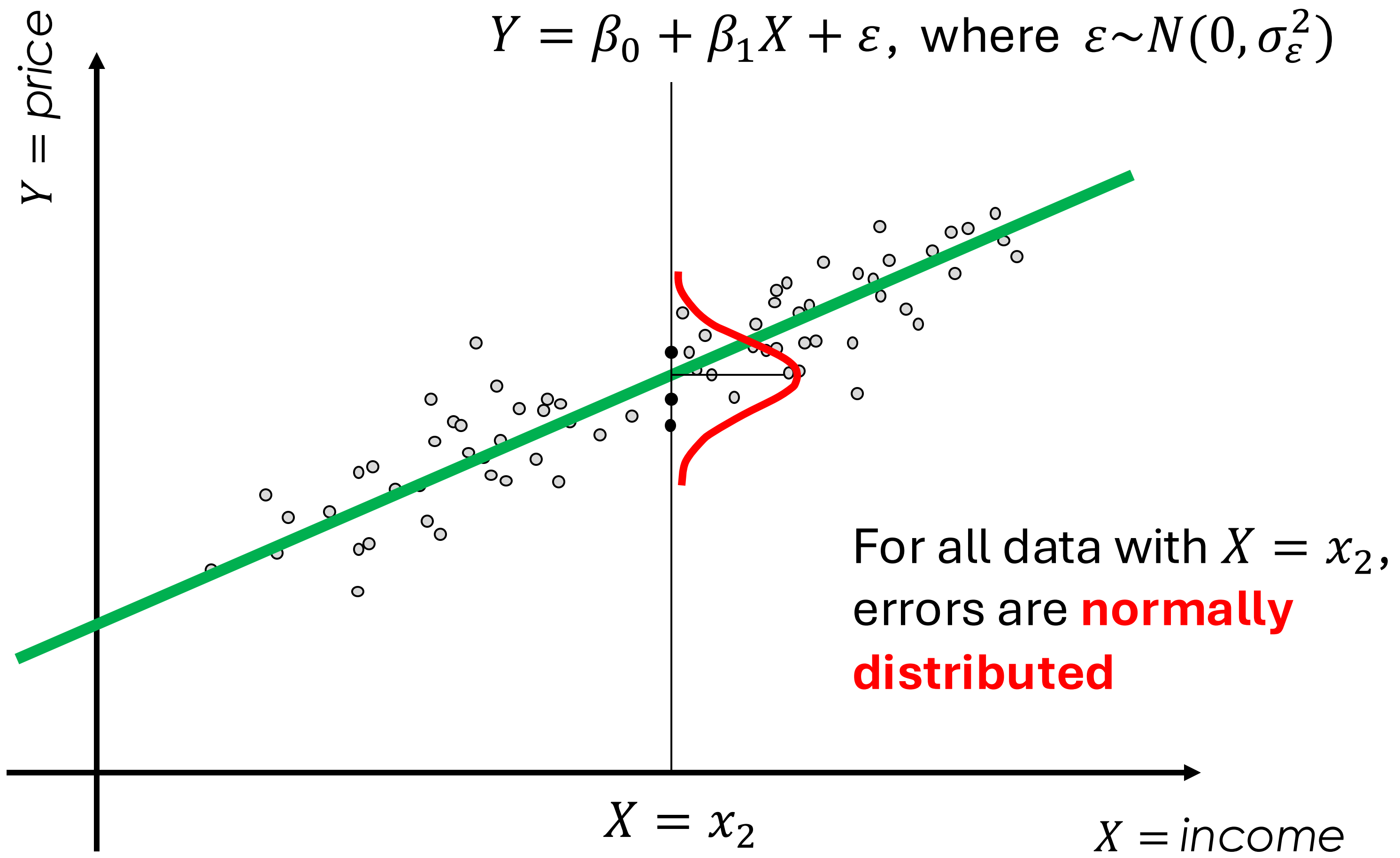
Simple Linear Regression: The Error



$Y = \beta_0 + \beta_1 X + \varepsilon,$ where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$

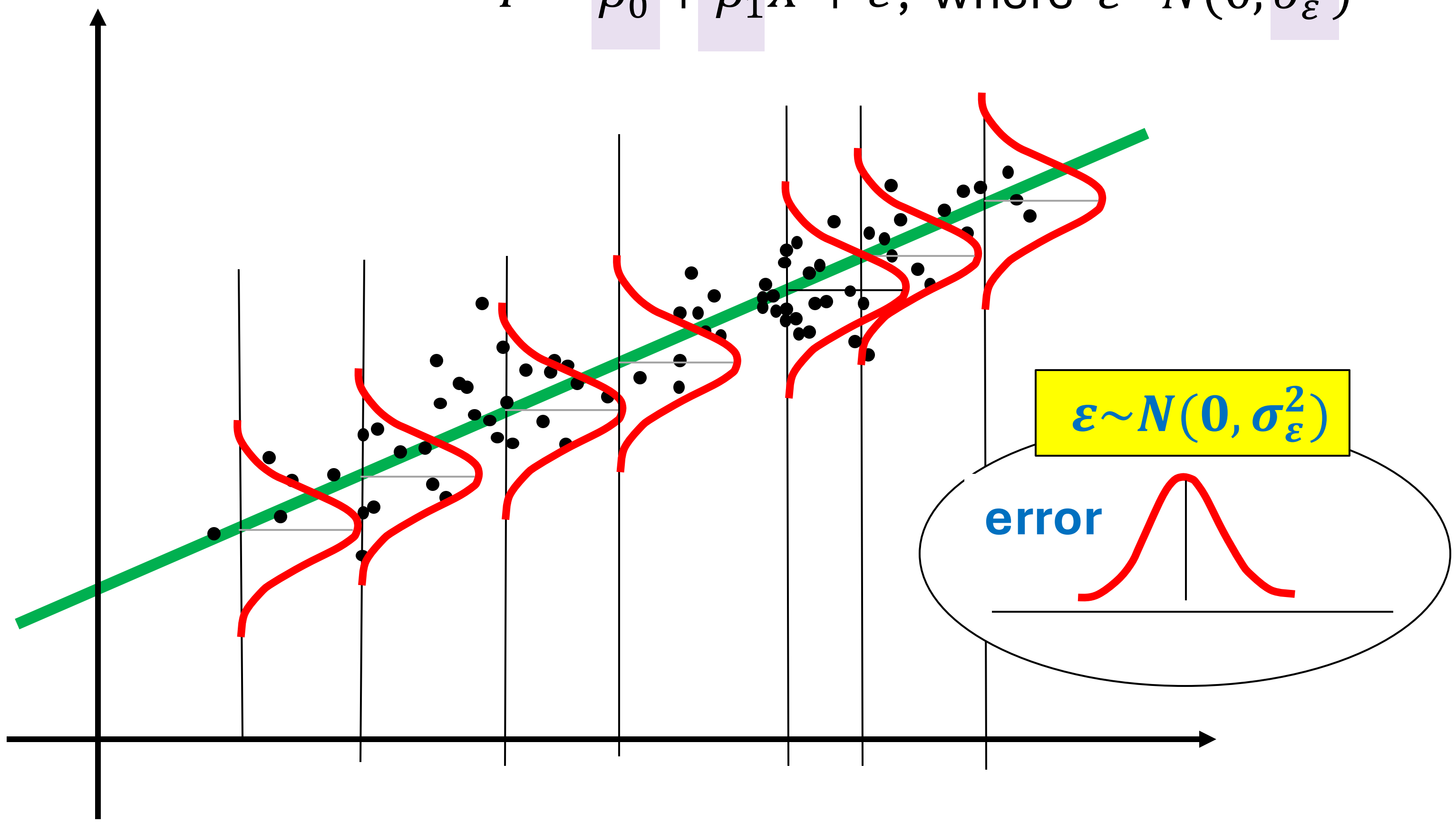
For all data with $X = x_1$, errors are **normally distributed**

Simple Linear Regression: The Error



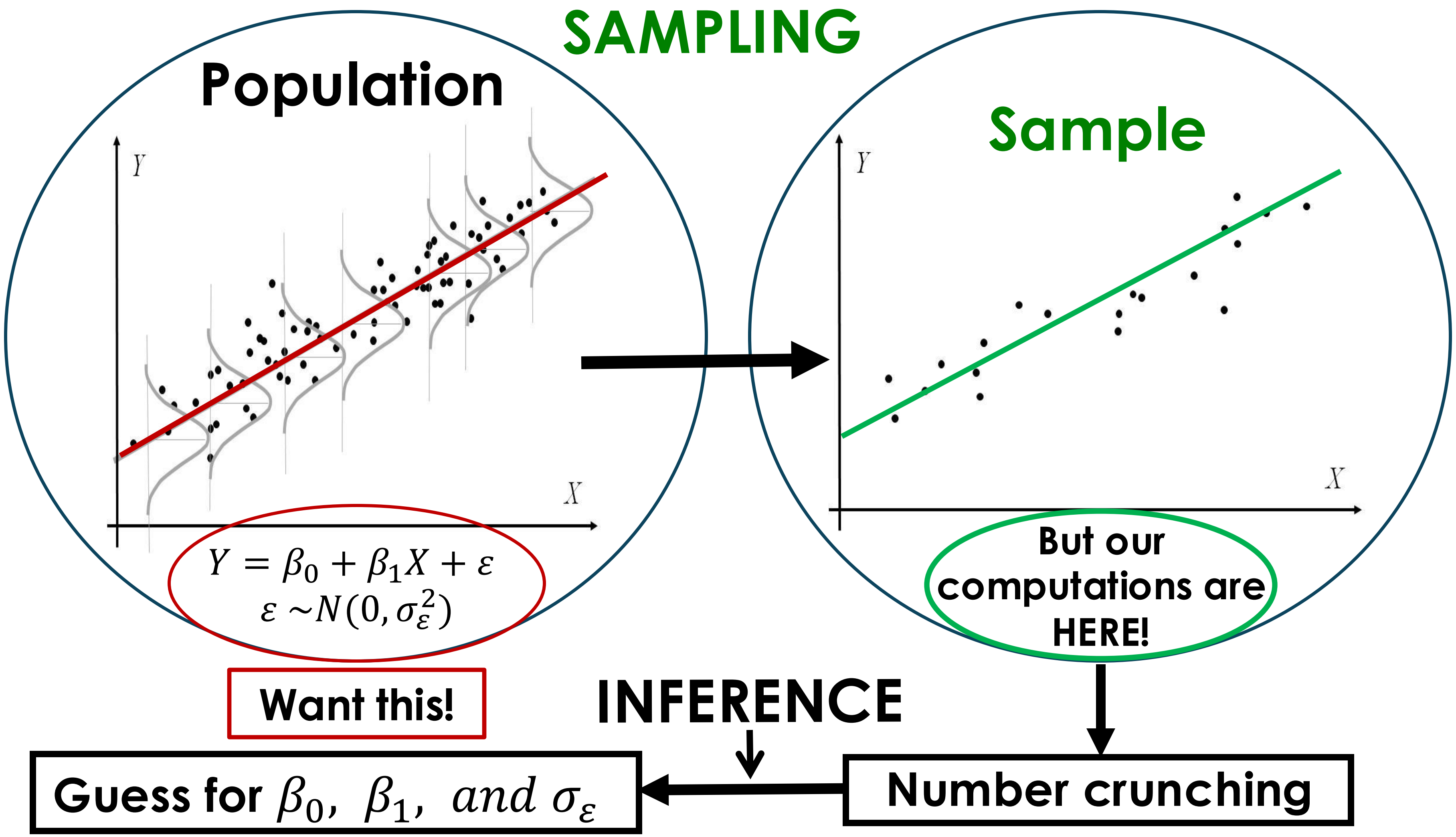
Simple Linear Regression: The Error

$Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$



The linear regression model is completely described by β_0, β_1 and σ_ε

Linear Regression: Population vs Sampling



Population Model vs Sample Model

- Regression equation for the ***population model***:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

- Regression equation for the ***sample model***:

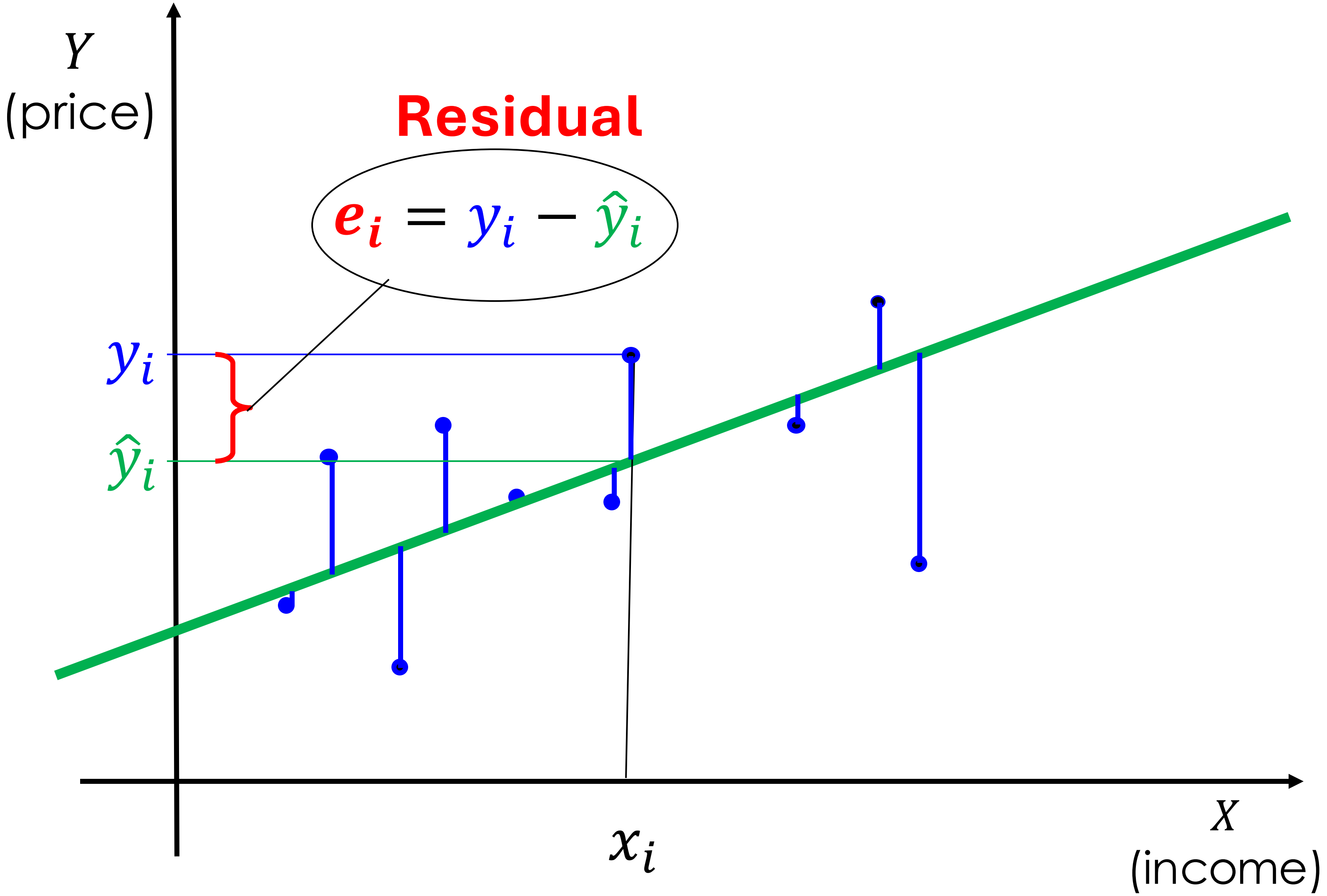
$$Y = \hat{\beta}_0 + \hat{\beta}_1 X, \qquad \varepsilon \sim N(0, S_\varepsilon^2)$$

Estimate these population parameters

$$\left\{ \begin{array}{lcl} \beta_0 & \longleftarrow & \hat{\beta}_0 \\ \beta_1 & \longleftarrow & \hat{\beta}_1 \\ \sigma_\varepsilon & \longleftarrow & S_\varepsilon \end{array} \right\}$$

Using these sample characteristics

Estimate The Regression Line



Estimate The Regression Line

Consider the i^{th} person

- y_i is the observed value of Y when $X = x_i$
- $\hat{y}_i$ is the estimated value of Y when $X = x_i$; $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residual: $e_i = y_i - \hat{y}_i,\ \ i = 1, \ldots, n$

Choose $\hat{\beta}_0, \hat{\beta}_1$ so that the sum squared residuals is as small as possible, i.e., minimize

$$SSE = \sum_i (e_i)^2 = \sum_i (y_i - \hat{y}_i)^2$$

Least Squares Method

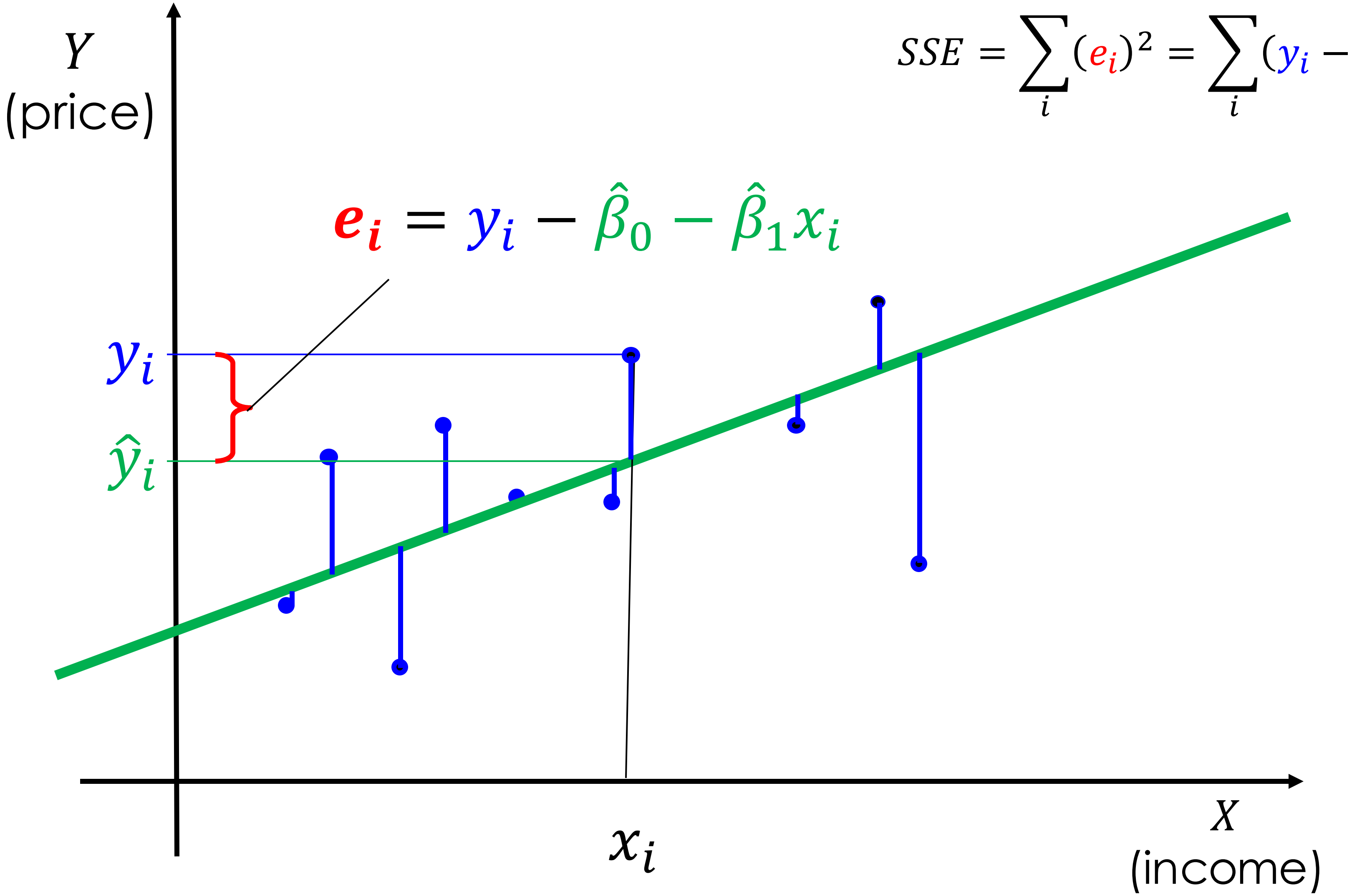
Wiggle the estimated regression line around until you find the minimum of SSE.

Python does it for us!

Estimate The Regression Line

Choose $\hat{\beta}_0, \hat{\beta}_1$ so that the sum squared residuals is as small as possible:

$$SSE = \sum_i (e_i)^2 = \sum_i (y_i - \hat{y}_i)^2$$



Focus 1: Estimate the Linear Relationship

- Obtain the regression line (Get $\hat{\beta}_0$, $\hat{\beta}_1$, S_ε)
- Point estimation of β_0 and β_1 (Interpret the regression line)

Example

Pittsburgh Housing Prices

Population Model vs. Sample Model

Regression equation for the ***population model***:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Regression equation for the ***sample model***:

$$Y = \widehat{\beta}_0 + \widehat{\beta}_1 X, \qquad \varepsilon \sim N(0, S_\varepsilon^2)$$

Estimate these population parameters

$$\left\{\begin{array}{l} \beta_0 \leftarrow \widehat{\beta}_0 \\ \beta_1 \leftarrow \widehat{\beta}_1 \\ \sigma_\varepsilon \leftarrow S_\varepsilon \end{array}\right\}$$

Using these sample characteristics

Example

Pittsburgh Housing Prices

```
housing_raw = pd.read_csv('pgh_housing_raw.csv', low_memory=False)
housing_raw.columns
```

```
import pandas as pd
import statsmodels.api as sm

# Prepare variables
X = sm.add_constant(housing['FINISHEDLIVINGAREA'])
y = housing['SALEPRICE']

# Run regression
model = sm.OLS(y, X).fit()

# Print results
print(model.summary())
```

Example

Pittsburgh Housing Prices

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$
$$= -44610 + 98.473\, X$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SALEPRICE   R-squared:                       0.143
Model:                            OLS   Adj. R-squared:                  0.143
Method:                 Least Squares   F-statistic:                 3.546e+04
Date:                Tue, 03 Dec 2024   Prob (F-statistic):               0.00
Time:                        12:30:39   Log-Likelihood:            -2.8852e+06
No. Observations:              212020   AIC:                         5.770e+06
Df Residuals:                  212018   BIC:                         5.770e+06
Df Model:                           1
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const            -4.461e+04    994.845    -44.837      0.000   -4.66e+04   -4.27e+04
FINISHEDLIVINGAREA  98.4736      0.523    188.300      0.000      97.449      99.499
==============================================================================
Omnibus:                   919112.701   Durbin-Watson:                   1.745
Prob(Omnibus):                  0.000   Jarque-Bera (JB):   10774647866709.926
Skew:                         132.135   Prob(JB):                         0.00
Kurtosis:                   34925.579   Cond. No.                     4.43e+03
==============================================================================
```

Y

$\hat{\beta}_0$

X

$\hat{\beta}_1$

Population Model vs. Sample Model

Regression equation for the ***population model***:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Regression equation for the ***sample model***:

$$Y = \widehat{\beta}_0 + \widehat{\beta}_1 X, \qquad \varepsilon \sim N(0, S_\varepsilon^2)$$

Estimate these population parameters

$$\left\{\begin{array}{l} \beta_0 \leftarrow \widehat{\beta}_0 \\ \beta_1 \leftarrow \widehat{\beta}_1 \\ \sigma_\varepsilon \leftarrow S_\varepsilon \end{array}\right\}$$

Using these sample characteristics

Example
Pittsburgh Housing Prices

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$
$$= -44610 + 98.473\ X$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SALEPRICE   R-squared:                       0.143
Model:                            OLS   Adj. R-squared:                  0.143
Method:                 Least Squares   F-statistic:                 3.546e+04
Date:                Tue, 03 Dec 2024   Prob (F-statistic):               0.00
Time:                        12:30:39   Log-Likelihood:            -2.8852e+06
No. Observations:              212020   AIC:                         5.770e+06
Df Residuals:                  212018   BIC:                         5.770e+06
Df Model:                           1
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const            -4.461e+04    994.845    -44.837      0.000   -4.66e+04   -4.27e+04
FINISHEDLIVINGAREA  98.4736      0.523    188.300      0.000      97.449      99.499
==============================================================================
Omnibus:                   919112.701   Durbin-Watson:                   1.745
Prob(Omnibus):                  0.000   Jarque-Bera (JB):   10774647866709.926
Skew:                         132.135   Prob(JB):                         0.00
Kurtosis:                   34925.579   Cond. No.                     4.43e+03
==============================================================================
```

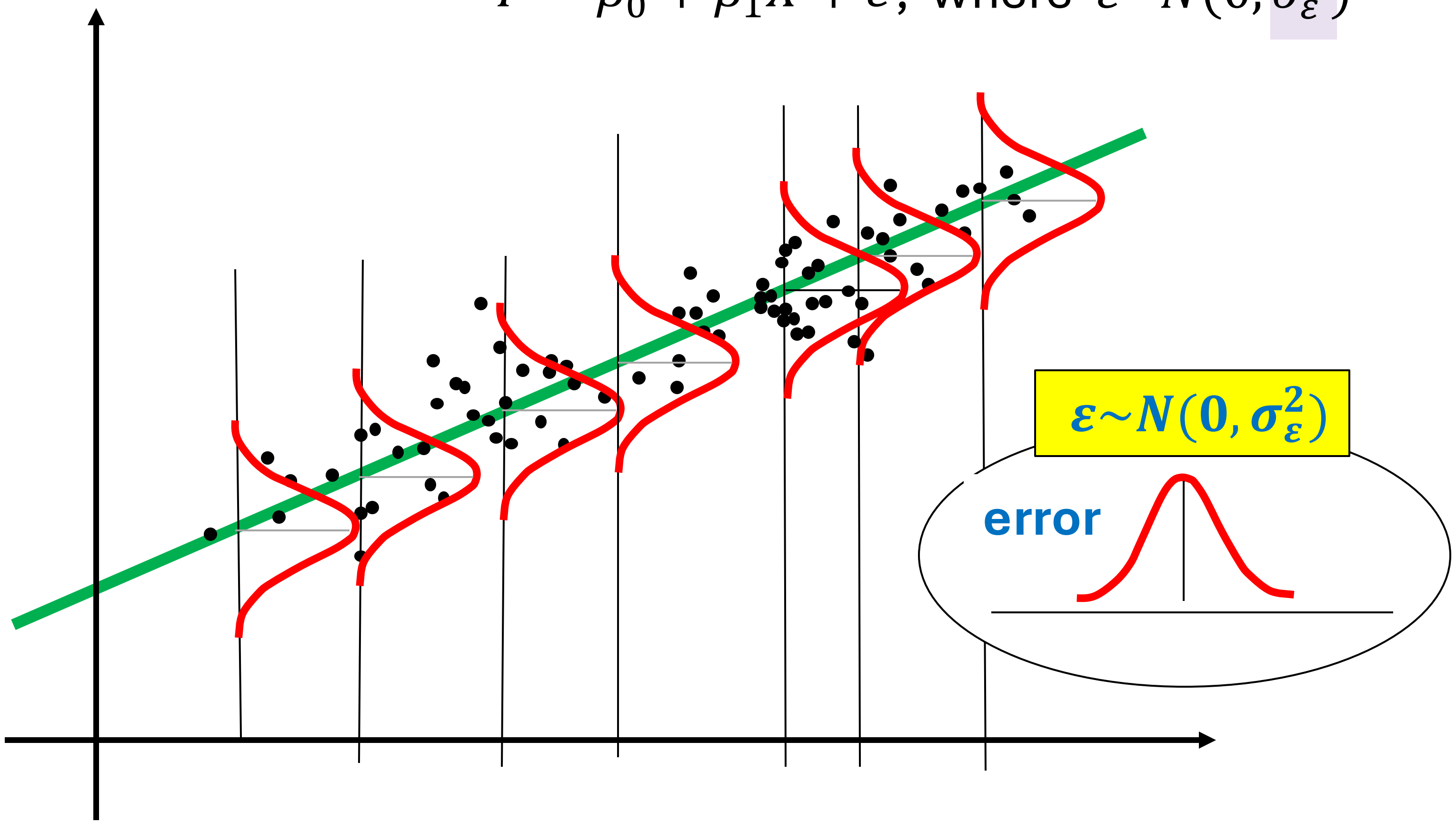
Y

X

$\hat{\beta}_0$

$\hat{\beta}_1$

$$Y = \beta_0 + \beta_1 X + \varepsilon, \text{ where } \varepsilon \sim N(0, \sigma_\varepsilon^2)$$



Example
Pittsburgh Housing Prices

```
np.sqrt(model.scale)
```

[184]:

```
196615.57246553455
```

Std. Err. Reg.

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$= -44610 + 98.473\, X$$

where $\varepsilon \sim N(0, S_\epsilon^2)$

where $\varepsilon \sim N(0, 196615^2)$

Standard Error of regression (S_ε)

- $\approx \sigma_\varepsilon$ in the distribution of the error term $\varepsilon \sim N(0, \sigma_\varepsilon^2)$
- Average deviation from the best fit line

Example

Pittsburgh Housing Prices

```
np.sqrt(model.scale)
```

[184]:

196615.57246553455

Std. Err. Reg.

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X \quad \text{where } \varepsilon \sim N(0, S_\epsilon^2)$$

$$= -44610 + 98.473\, X \quad \text{where } \varepsilon \sim N(0, 197.3^2)$$

$$Price = -44610 + 98.473\, Area, \qquad \varepsilon \sim N(0, 196615^2)$$

Standard Error of regression (S_ε)

- $\approx \sigma_\varepsilon$ in the distribution of the error term $\varepsilon \sim N(0, \sigma_\varepsilon^2)$
- Average deviation from the best fit line

Summary

Simple regression

- Y: dependent variable; X: independent variable
- Population model: $Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$
- Sample model: $Y = \hat{\beta}_0 + \hat{\beta}_1 X, \qquad \varepsilon \sim N(0, S_\varepsilon^2)$

Regression output

- Standard error of the regression (S_ε as the estimate of σ_ε)
- for the intercept β_0 and the slope β_1: $\hat{\beta}_j$

Focus 1: Estimate the Linear Relationship

- Obtain the regression line (Get $\hat{\beta}_0$, $\hat{\beta}_1$, S_ε)
- Point estimation of β_0 and β_1 (Interpret the regression line)
- Interval estimation of β_0 and β_1
- Inferences about the slope β_1

$$H_0\colon \beta_1 = 0$$

$$H_0\colon \beta_1 = \text{some number other than } 0$$

Example
Pittsburgh Housing Prices

$Price = -44.9781 + 0.0997\, Area, \varepsilon \sim N(0, 197.3^2)$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SALEPRICE   R-squared:                       0.143
Model:                            OLS   Adj. R-squared:                  0.143
Method:                 Least Squares   F-statistic:                 3.546e+04
Date:                Tue, 03 Dec 2024   Prob (F-statistic):               0.00
Time:                        12:30:39   Log-Likelihood:            -2.8852e+06
No. Observations:              212020   AIC:                         5.770e+06
Df Residuals:                  212018   BIC:                         5.770e+06
Df Model:                           1
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const              -4.461e+04    994.845    -44.837      0.000   -4.66e+04   -4.27e+04
FINISHEDLIVINGAREA    98.4736      0.523    188.300      0.000      97.449      99.499
======================================================================================
Omnibus:                   919112.701   Durbin-Watson:                   1.745
Prob(Omnibus):                  0.000   Jarque-Bera (JB):   10774647866709.926
Skew:                         132.135   Prob(JB):                         0.00
Kurtosis:                   34925.579   Cond. No.                     4.43e+03
==============================================================================
```

$\hat{\beta}_1$: point estimate of β_1

$S_{\hat{\beta}_1}$: std. dev. (std. error) of $\hat{\beta}_1$

units of $S_{\hat{\beta}_1}$ that $\hat{\beta}_1$ is from 0

p-value to test $H_0 : \beta_1 = 0$

I p < 0.05, $\beta_1 \neq 0$ at 95%

95% C.I for β_1

H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$

Summary

- Standard Error of $\hat{\beta}_j$
- Confident interval of β_j
- Hypothesis testing on the slope coefficient β_1

$$H_0: \beta_1 = 0$$

$$H_0: \beta_1 = \text{some number other than } 0$$

Focus 2: How good is the model fit?

- Does the relationship between X and Y follow:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

- Does the population model (population relationship) exist at all?
- There are different ways to answer this question
  1. Test statistics, *p*-value of the regression coefficient
  2. Standard Error of the Regression
  3. R^2 (Coefficient of Determination)

Example

Pittsburgh Housing Prices

$Price = -44.9781 + 0.0997\, Area, \varepsilon \sim N(0, 197.3^2)$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SALEPRICE   R-squared:                       0.143
Model:                            OLS   Adj. R-squared:                  0.143
Method:                 Least Squares   F-statistic:                 3.546e+04
Date:                Tue, 03 Dec 2024   Prob (F-statistic):               0.00
Time:                        12:30:39   Log-Likelihood:            -2.8852e+06
No. Observations:              212020   AIC:                         5.770e+06
Df Residuals:                  212018   BIC:                         5.770e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const            -4.461e+04    994.845    -44.837      0.000   -4.66e+04   -4.27e+04
FINISHEDLIVINGAREA  98.4736      0.523    188.300      0.000      97.449      99.499
==============================================================================
Omnibus:               919112.701   Durbin-Watson:                   1.745
Prob(Omnibus):              0.000   Jarque-Bera (JB):   10774647866709.926
Skew:                     132.135   Prob(JB):                         0.00
Kurtosis:               34925.579   Cond. No.                     4.43e+03
==============================================================================
```

R^2: Coefficient of Determination

Proportion of the variation in Y explained by the regression model

$\beta_1 \neq 0$, linear relationship exists

What is R^2?

Proportion of the variation in Y that can be explained by the regression model:

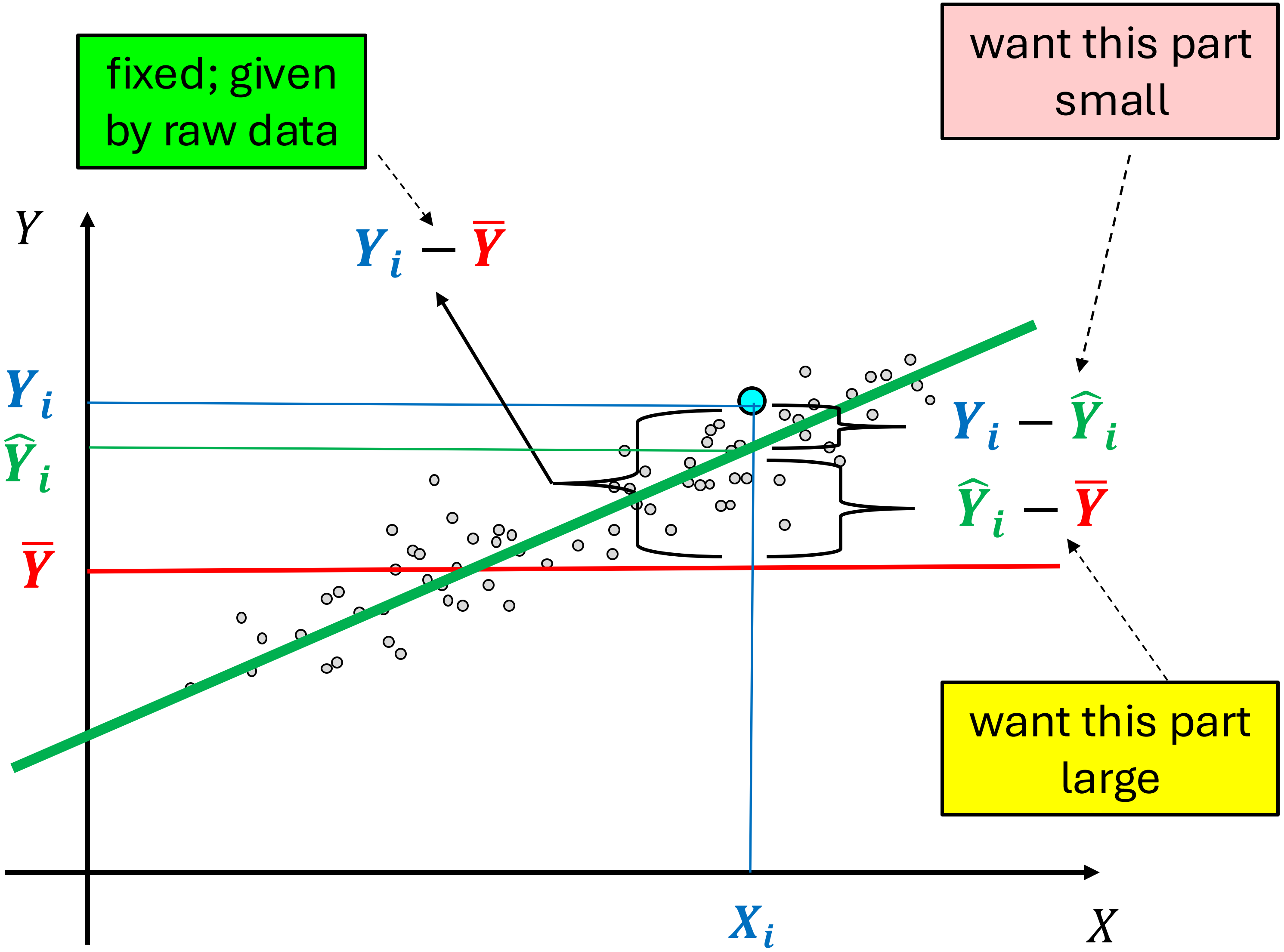
$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

Proportion of the variation in Y that can be explained by:

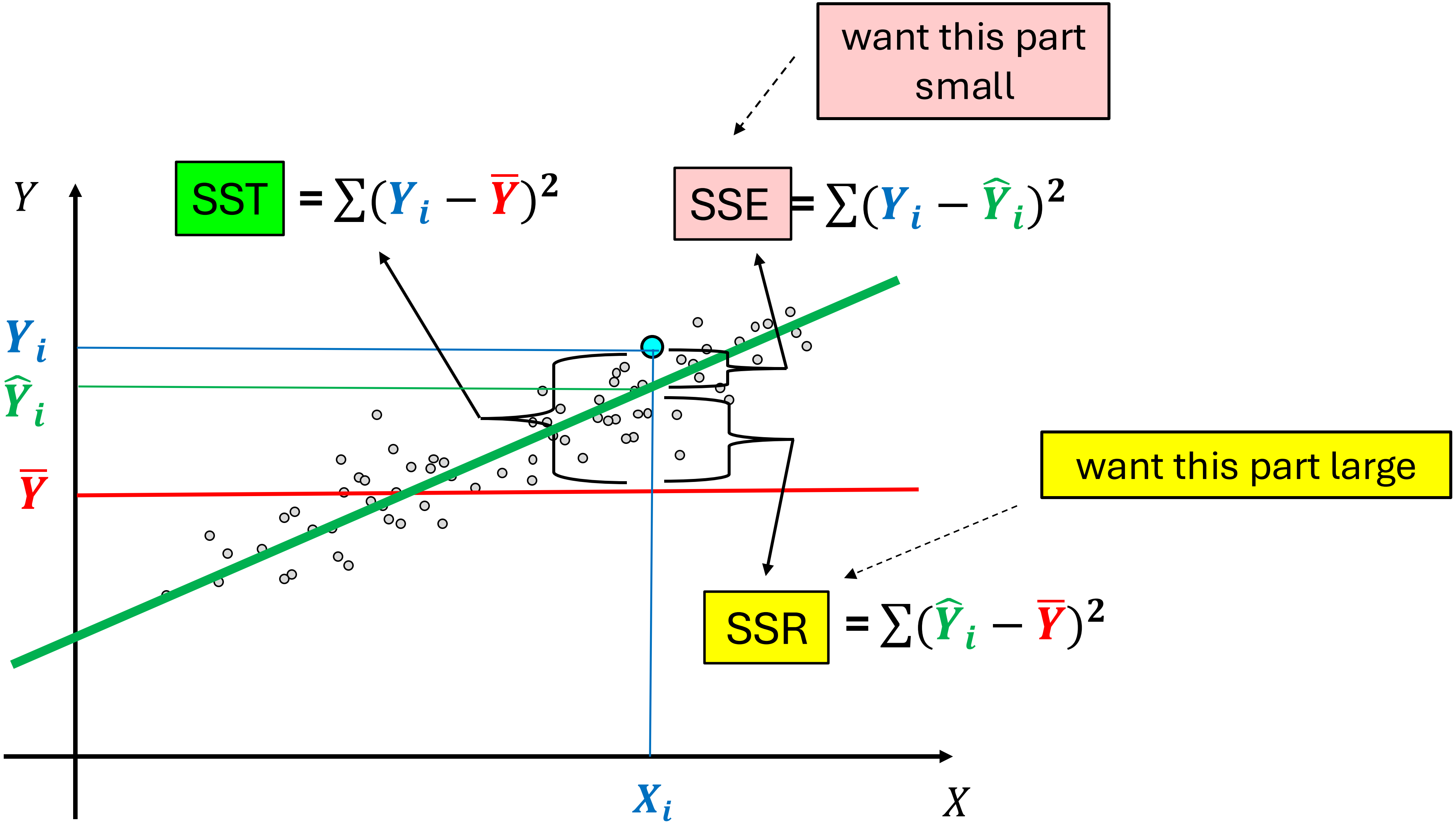
(1) the variation in X

(2) the estimated relationship $Y = \hat{\beta}_0 + \hat{\beta}_1 X$

Measures of Variation



Measures of Variation



Measures of Variation

Sum Squared Total

$\boxed{\text{SST}} = \sum(Y_i - \bar{Y})^2$

Fixed by data

- Measures total sample variation of Y around its mean $\bar{Y}$

Sum Squared Regression

$\boxed{\text{SSR}} = \sum(\hat{Y}_i - \bar{Y})^2$

want this large

- Measures variation in Y attributable to the regression line (i.e. factors that can be explained by X and the regression line)

Sum Squared Error

$\boxed{\text{SSE}} = \sum(Y_i - \hat{Y}_i)^2$

want this small

- Measures variation in Y attributable to factors other than X (i.e., factors not captured by X and the regression line)

What is R^2?

… the proportion of the variation in Y in the sample that can be explained by the regression model.

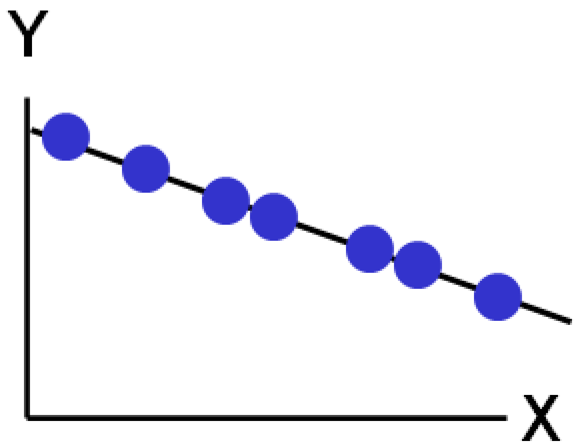
$$\boxed{R^2} = \frac{\boxed{\text{SSR}}}{\boxed{\text{SST}}}$$

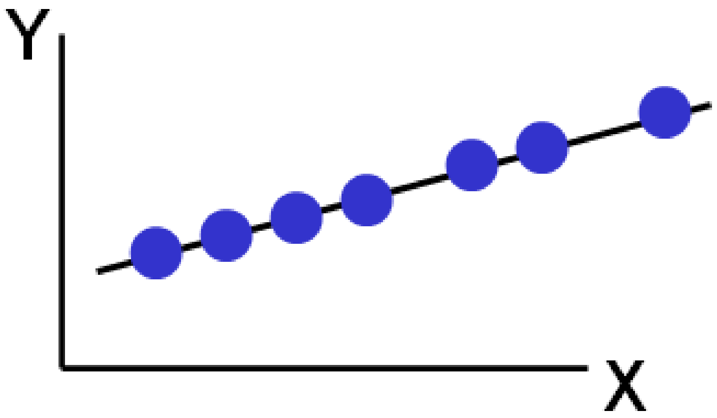
$$0 \leq R^2 \leq 1$$

What is R^2?

… the proportion of the variation in Y in the sample that can be explained by the regression model.

$\boxed{R^2 = 1}$





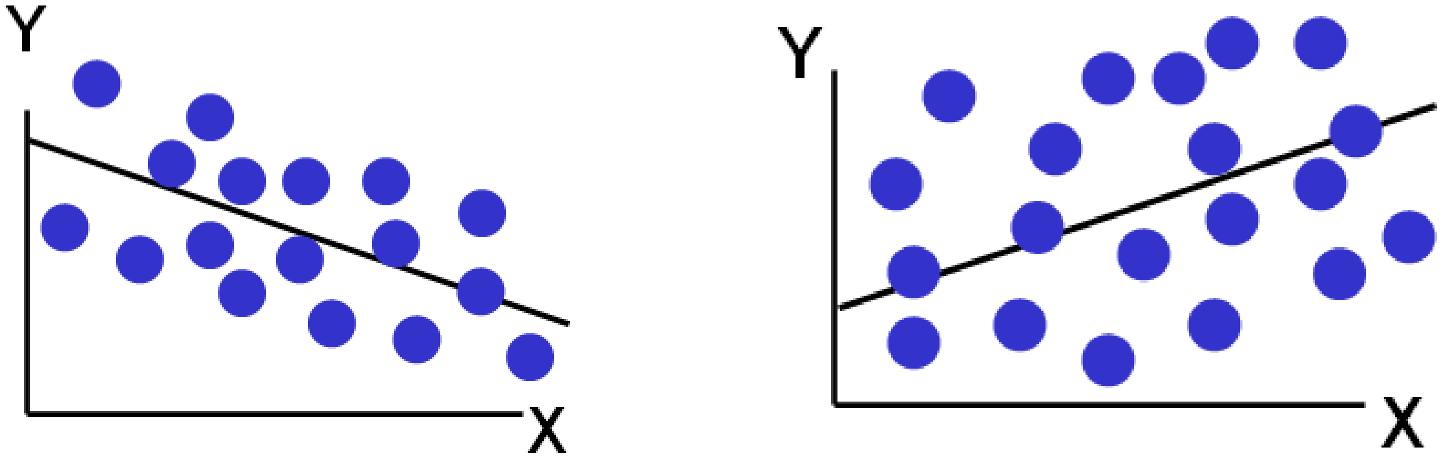
Perfect linear relationship between X and Y:

- 100% of the variation in Y is explained by variation in X

What is R^2?

… the proportion of the variation in Y in the sample that can be explained by the regression model.

$$0 < R^2 < 1$$



Weaker linear relationship between X and Y:

- Some but not all the variation in Y is explained by variation in X

But be careful…

R^2 has a very appealing interpretation but…

- Often meaningless to compare R^2 between models.
- Might provide no info on the prediction power and inferential quality of the model
- Could be more worthwhile to explain 50% variation of Y in Model 1 than to explain 70% variation of Y in Model 2.

Always think about the goal of your analysis:

- What are you using a regression for?

Summary

Simple regression

- Y: dependent variable; X: independent variable
- Population model: $Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$
- Sample model: $\quad Y = \hat{\beta}_0 + \hat{\beta}_1 X, \qquad \varepsilon \sim N(0, S_\varepsilon^2)$

- Regression output
  - Standard error of the regression: S_ε
  - Coefficient of determination: R^2
  - Point & interval estimations, and hypothesis testing of $\hat{\beta}_j$