Linear regression

Fourth Session

Orlando Sabogal-Cardona
PhD researcher
University College London UCL

The basic problem: fit a line

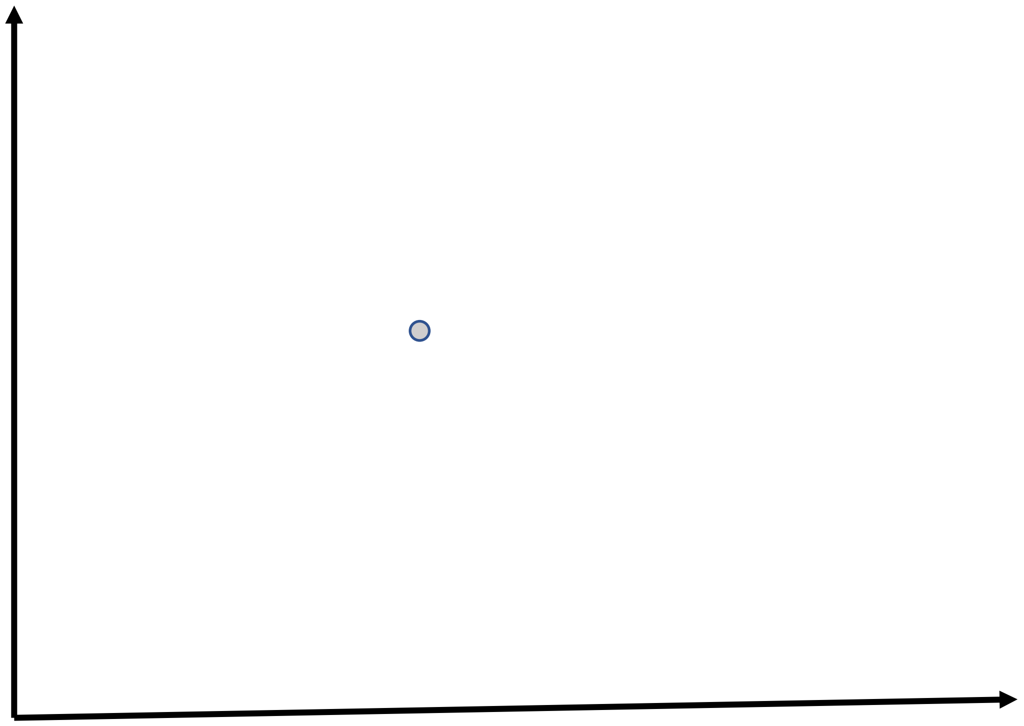
Y	X
Y_1	X_1
Y_2	X_2
Y_3	X_3
$\vdots$	$\vdots$
Y_n	X_n

The basic problem: fit a line

Y	X
Y_1	X_1
Y_2	X_2
Y_3	X_3
$\vdots$	$\vdots$
Y_n	X_n

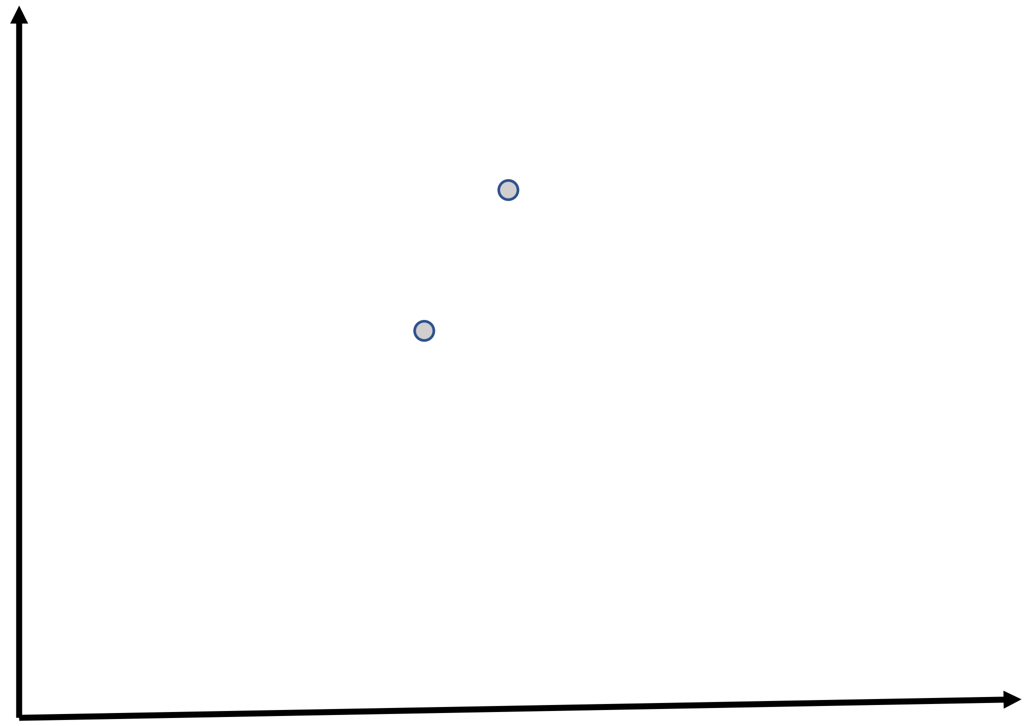
The basic problem: fit a line

Y	X
Y_1	X_1
Y_2	X_2
Y_3	X_3
$\vdots$	$\vdots$
Y_n	X_n



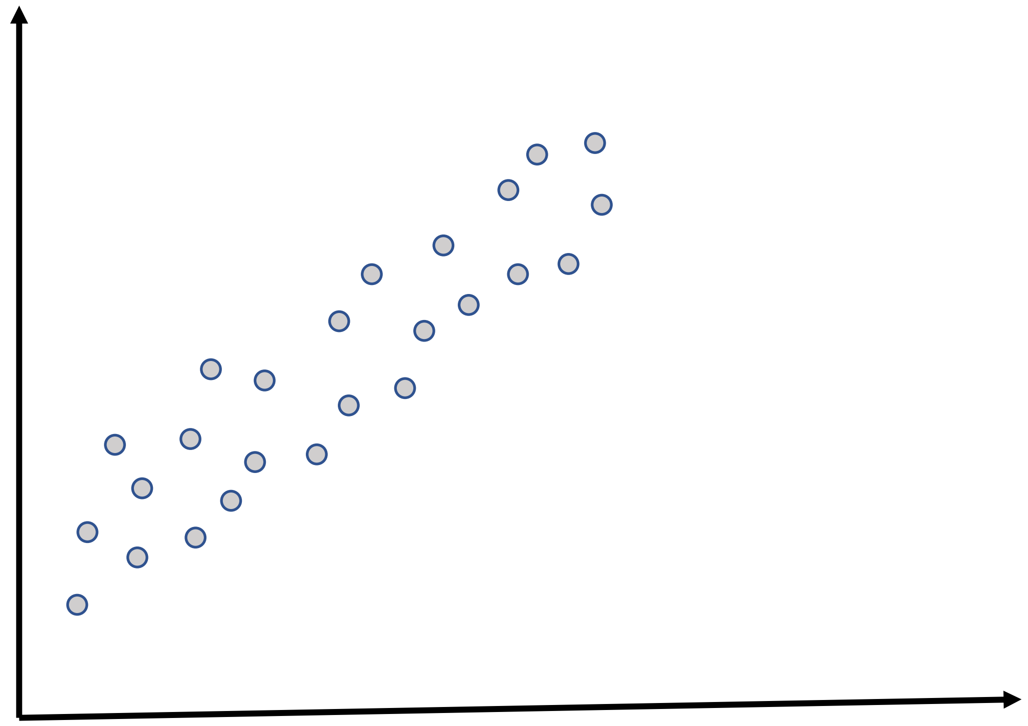
The basic problem: fit a line

Y	X
Y_1	X_1
Y_2	X_2
Y_3	X_3
$\vdots$	$\vdots$
Y_n	X_n

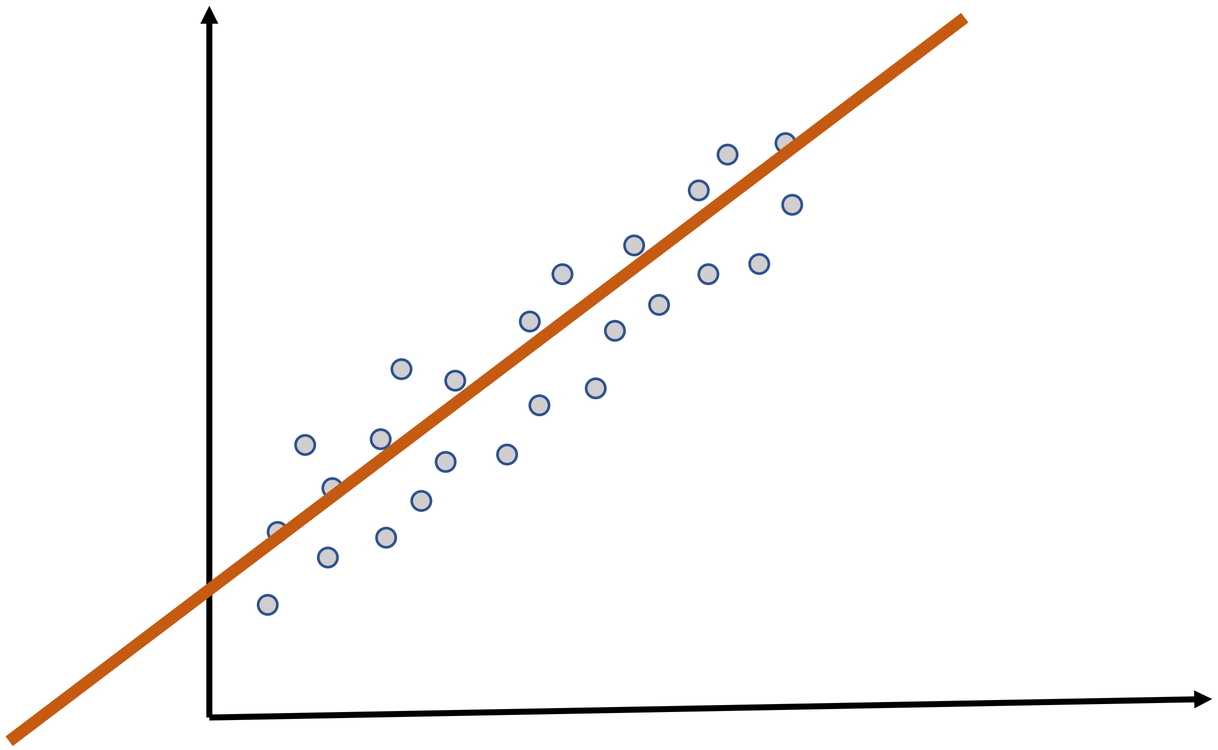


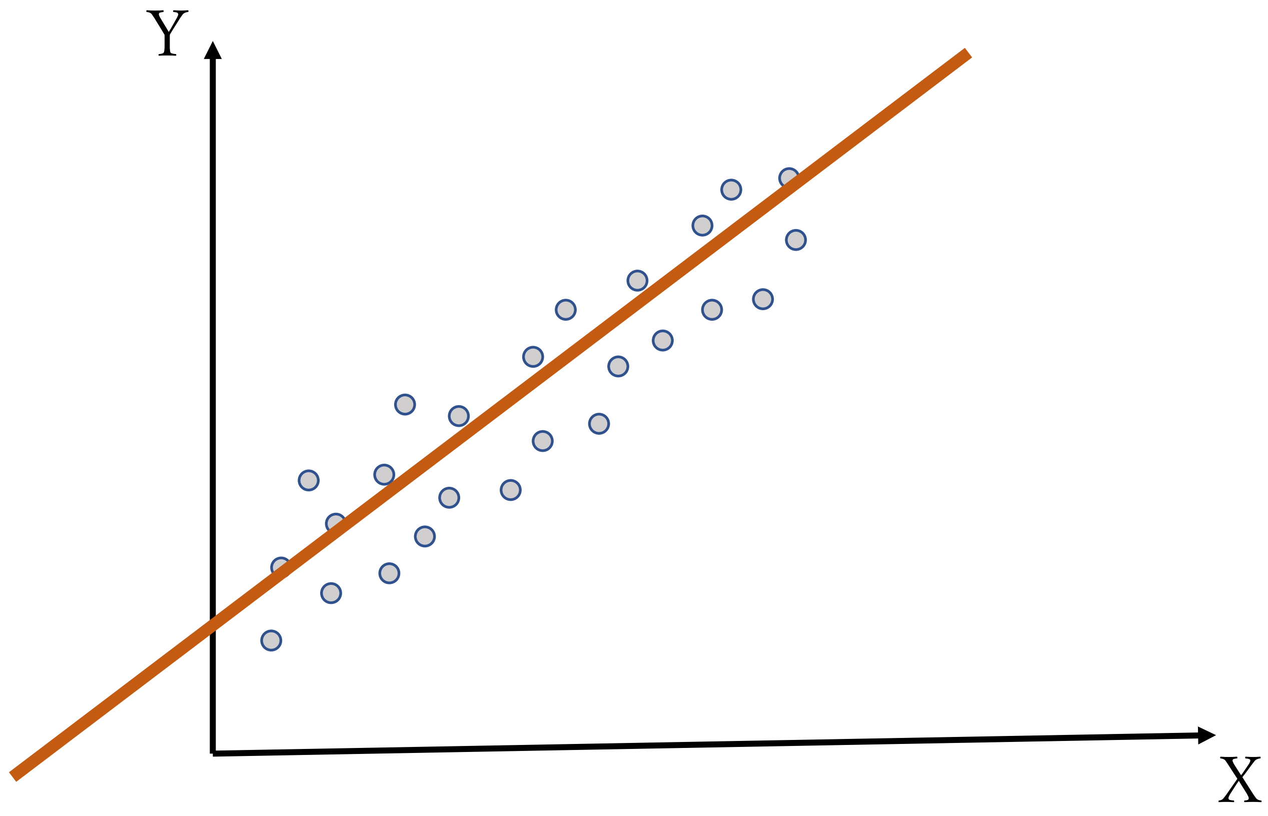
The basic problem: fit a line

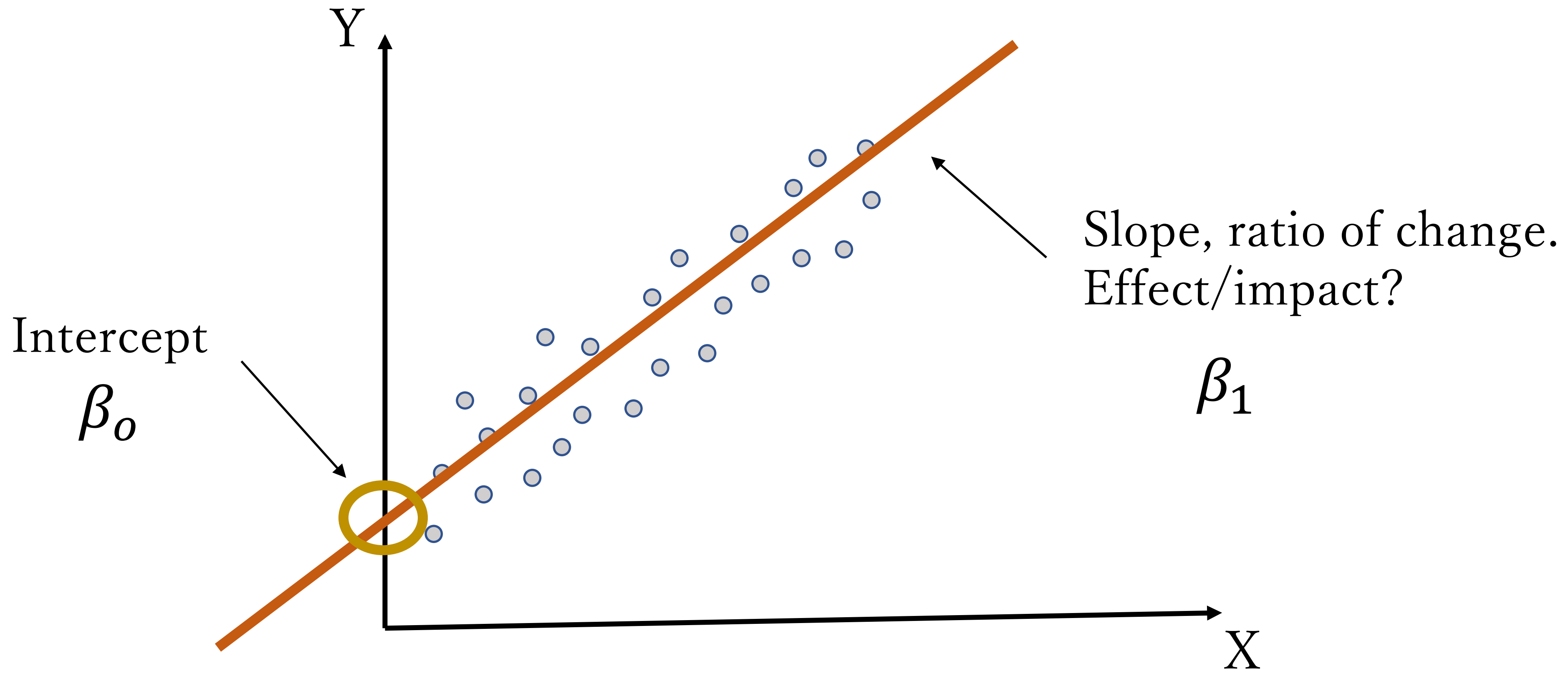
Y	X
Y_1	X_1
Y_2	X_2
Y_3	X_3
$\vdots$	$\vdots$
Y_n	X_n

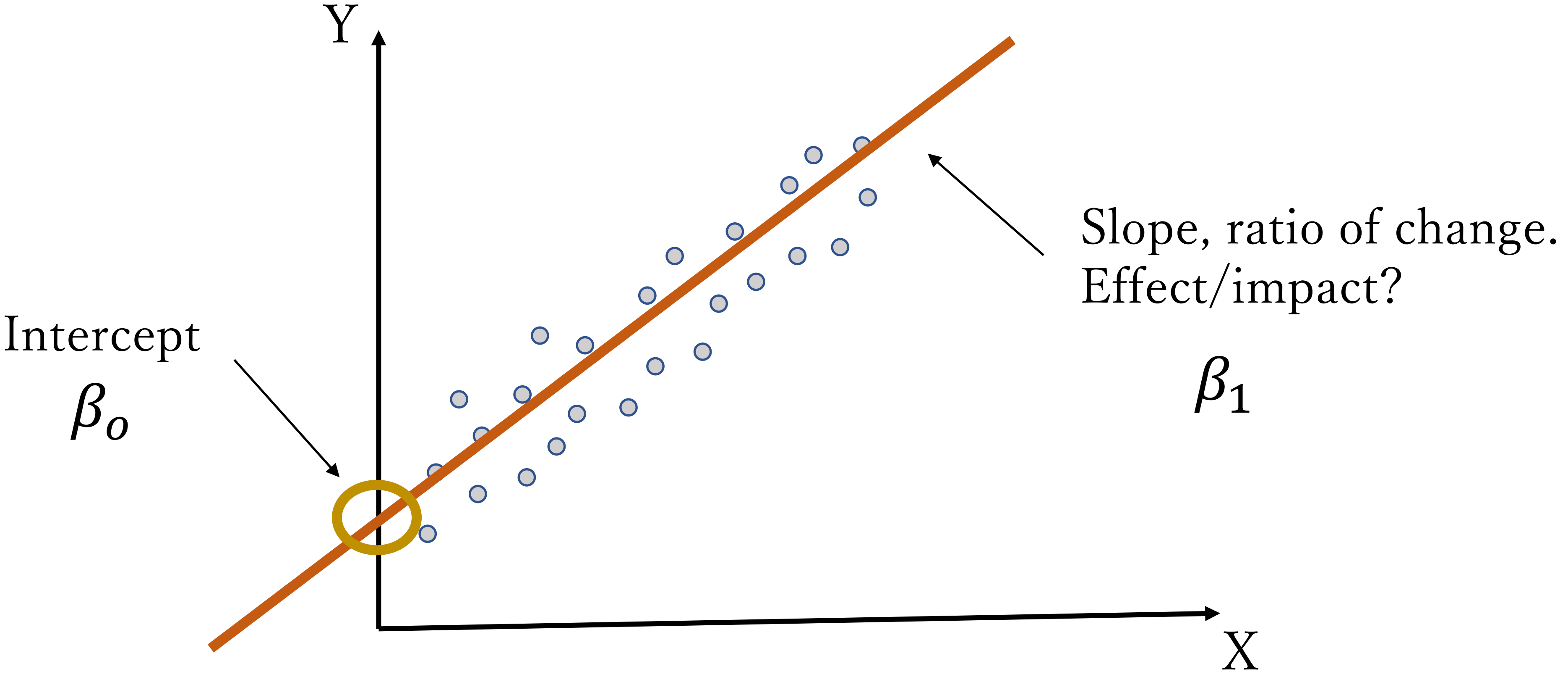


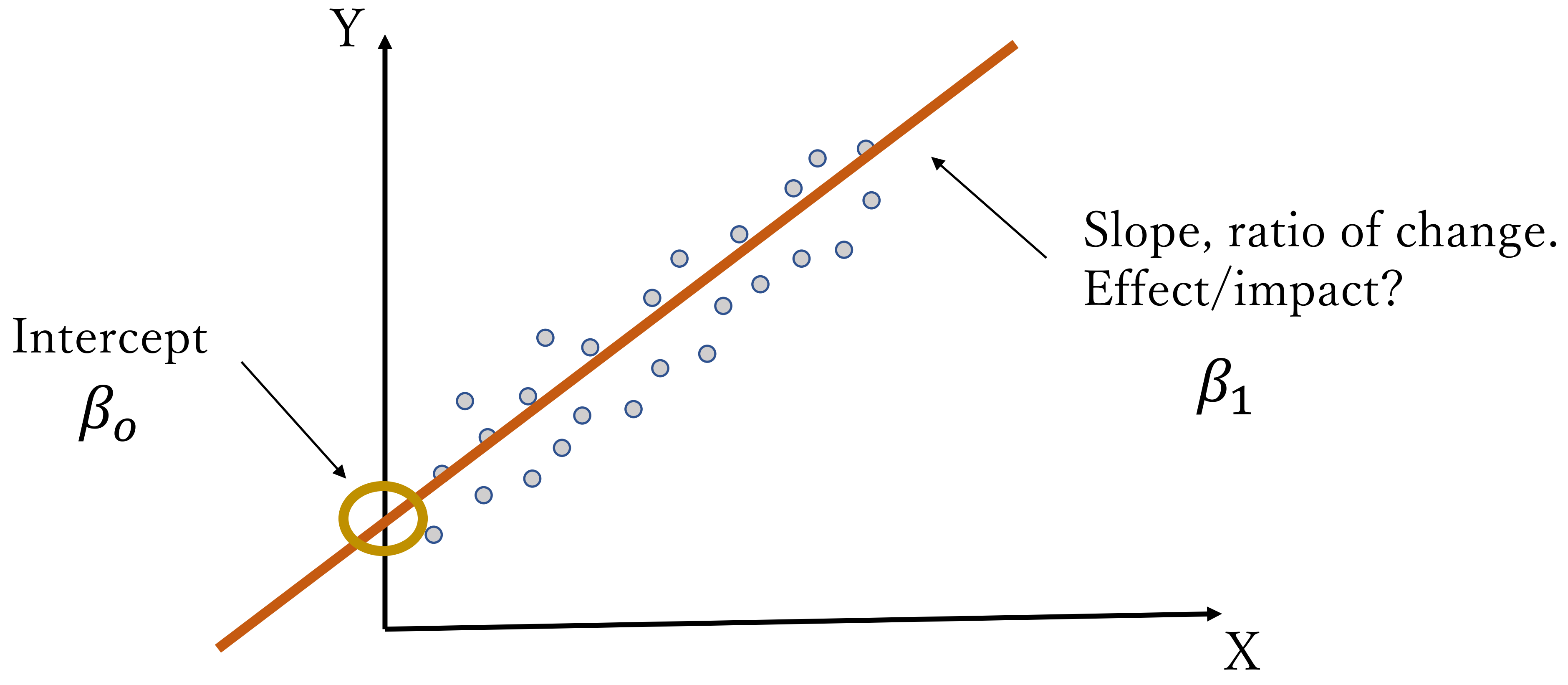
The basic problem: fit a line



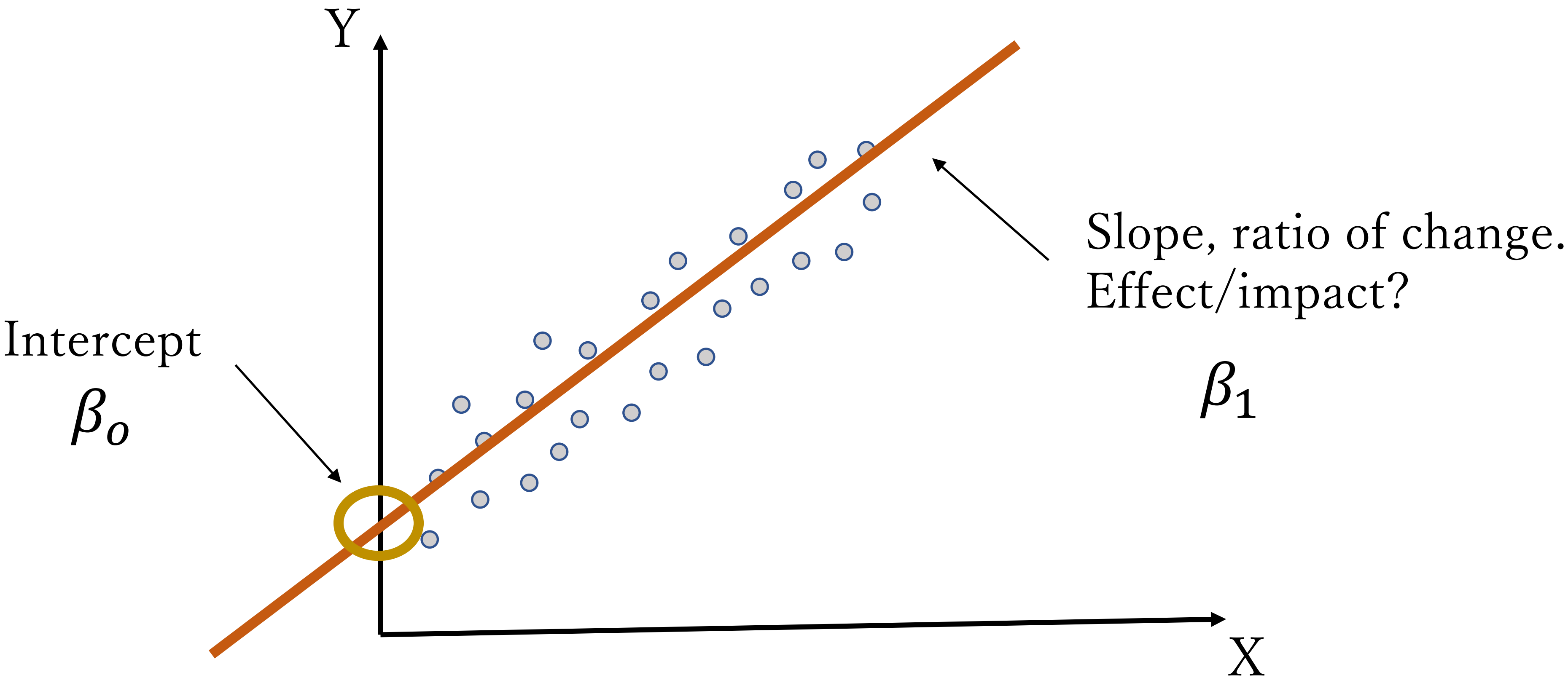

Y
X


Y
X
Intercept
β_o
Slope, ratio of change.
Effect/impact?
β_1


Y
X
Intercept
β_o
Slope, ratio of change.
Effect/impact?
β_1



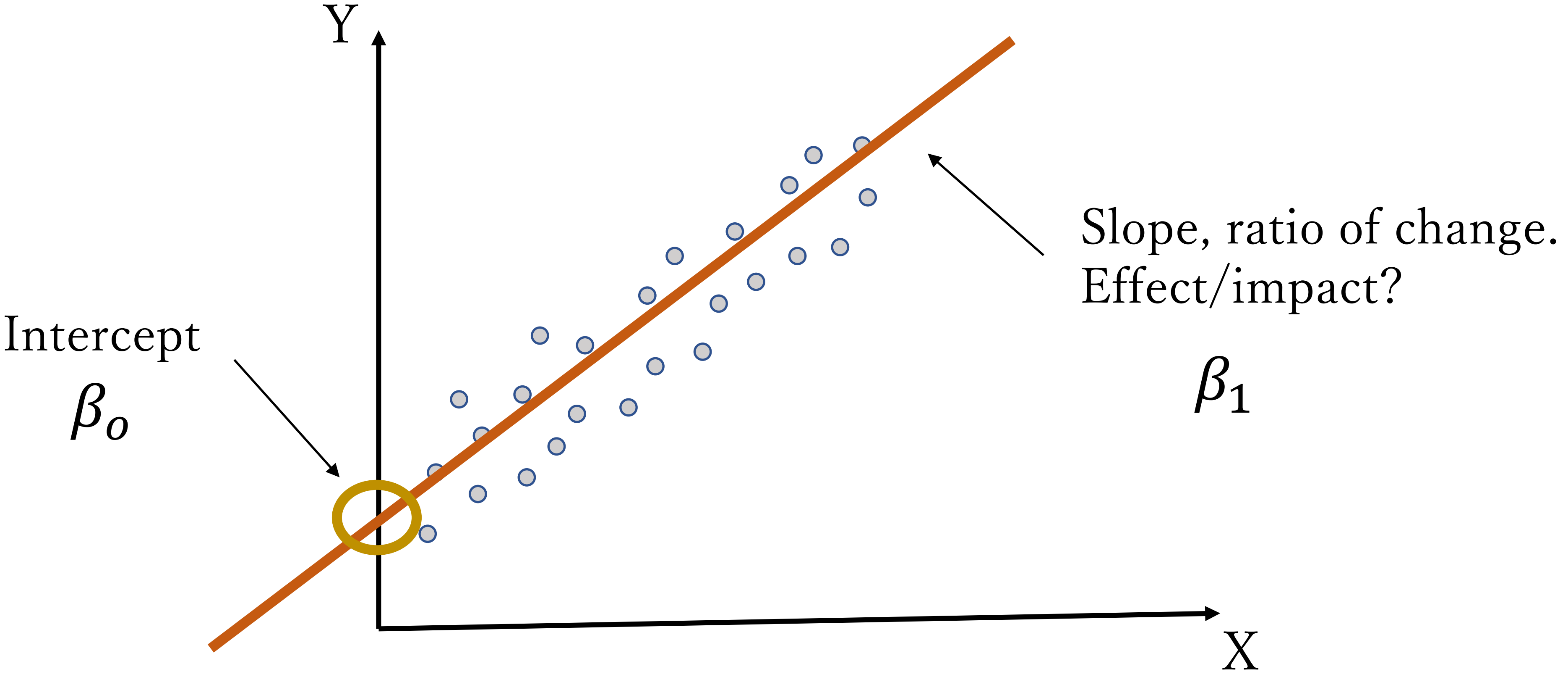
$$Y = \beta_0 + \beta_1 X$$



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

But remember: we have a sample and we do not know the parameters β_0 and β_1

A useful way to think about this equation is a the "data generator process"

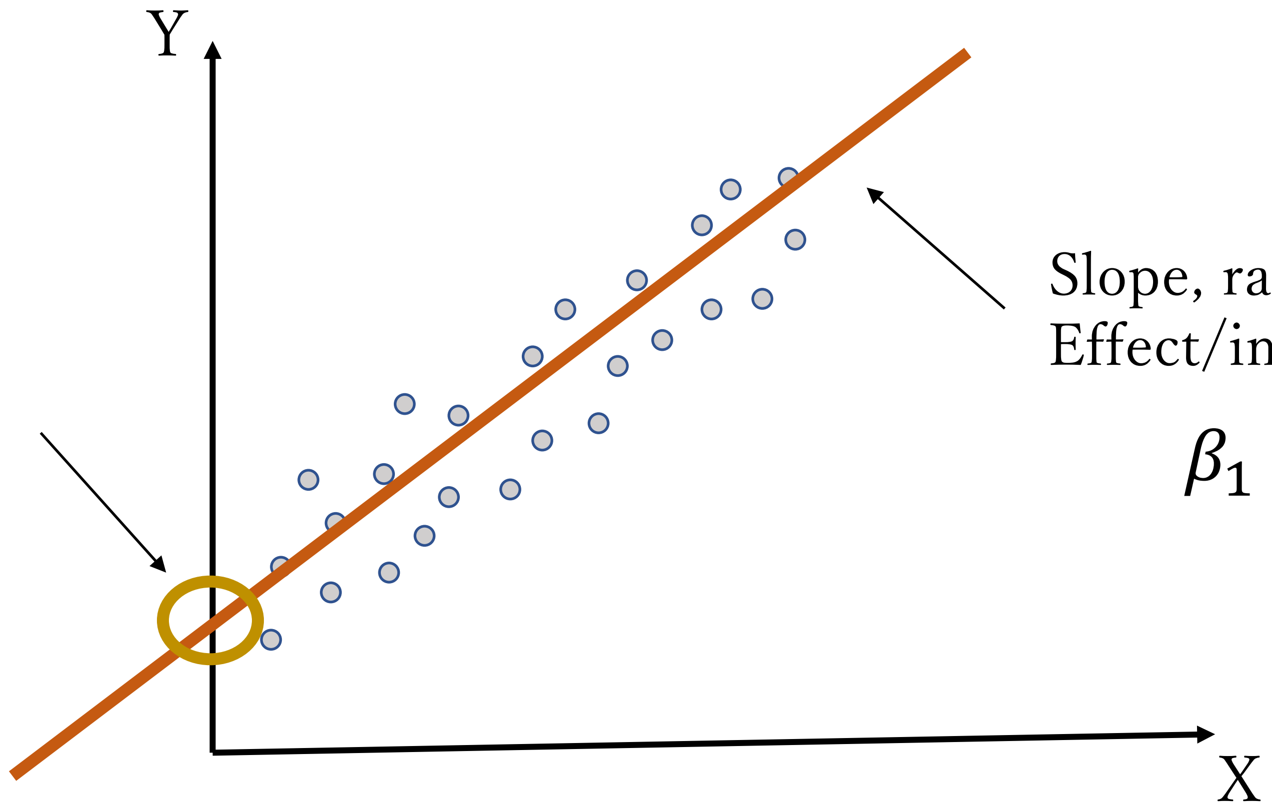


$$Y = \beta_0 + \beta_1 X$$

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X + e$$

Intercept

β_o



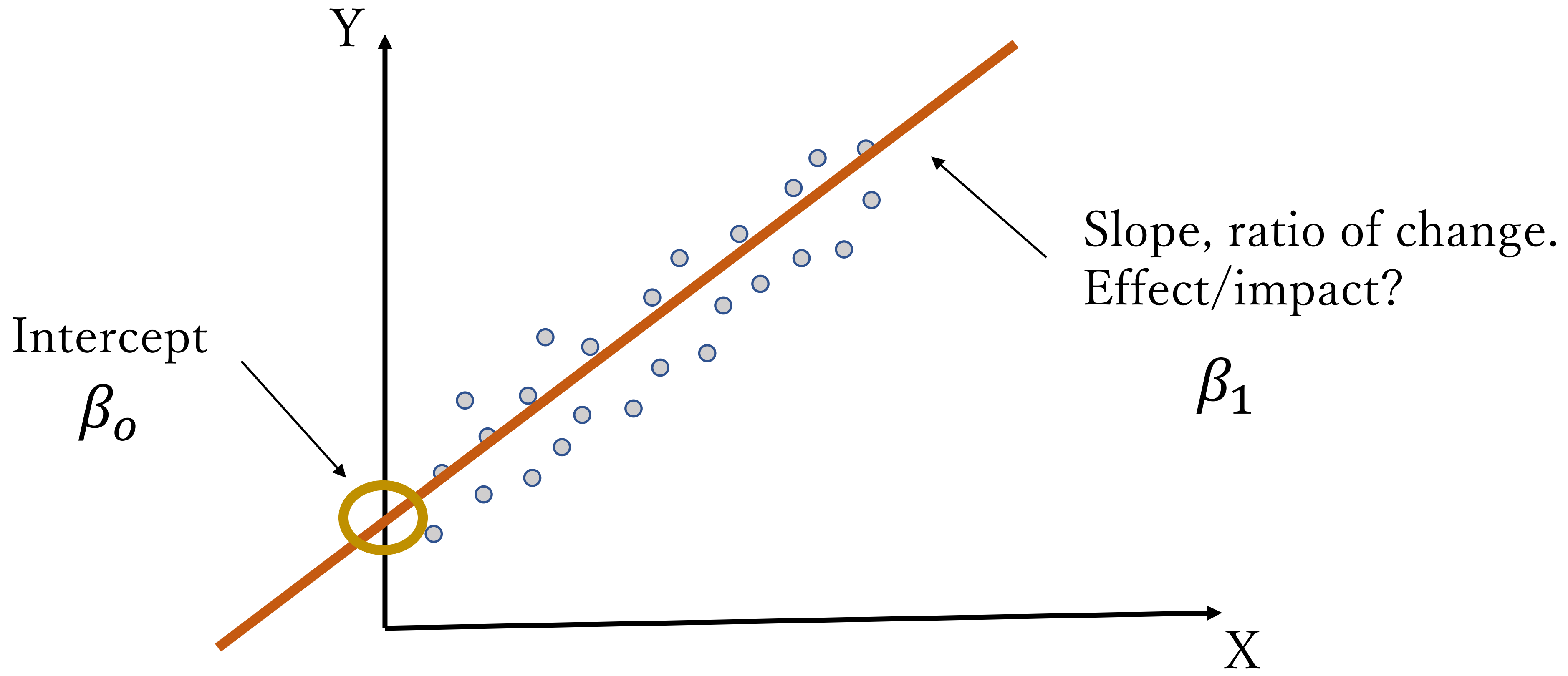
Slope, ratio of change.
Effect/impact?

β_1

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

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X + e$$

Take a minute here to remember the Central Limit Theorem



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

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X + e$$

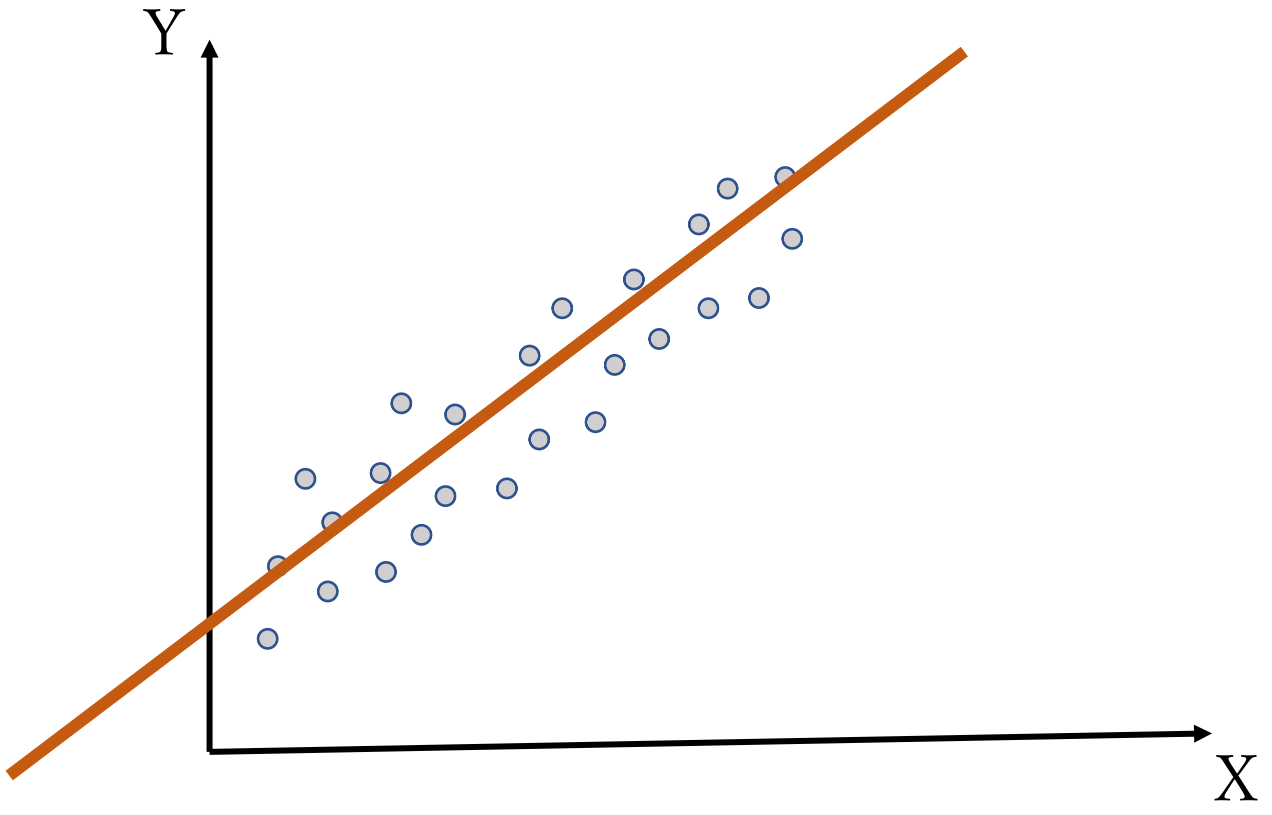
$$e = Y - \hat{Y}$$ → Error/residual

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

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X + e$$

$$e = Y - \hat{Y}$$

We need the parameters (β_0 and β_1) that minimize the residuals of all observations



$$min \sum_{i=1}^{n} e^2$$

$$min \sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} + \widehat{\beta_1} X_i\right)^2$$

Sum of the squared residuals
Why not only the residuals?

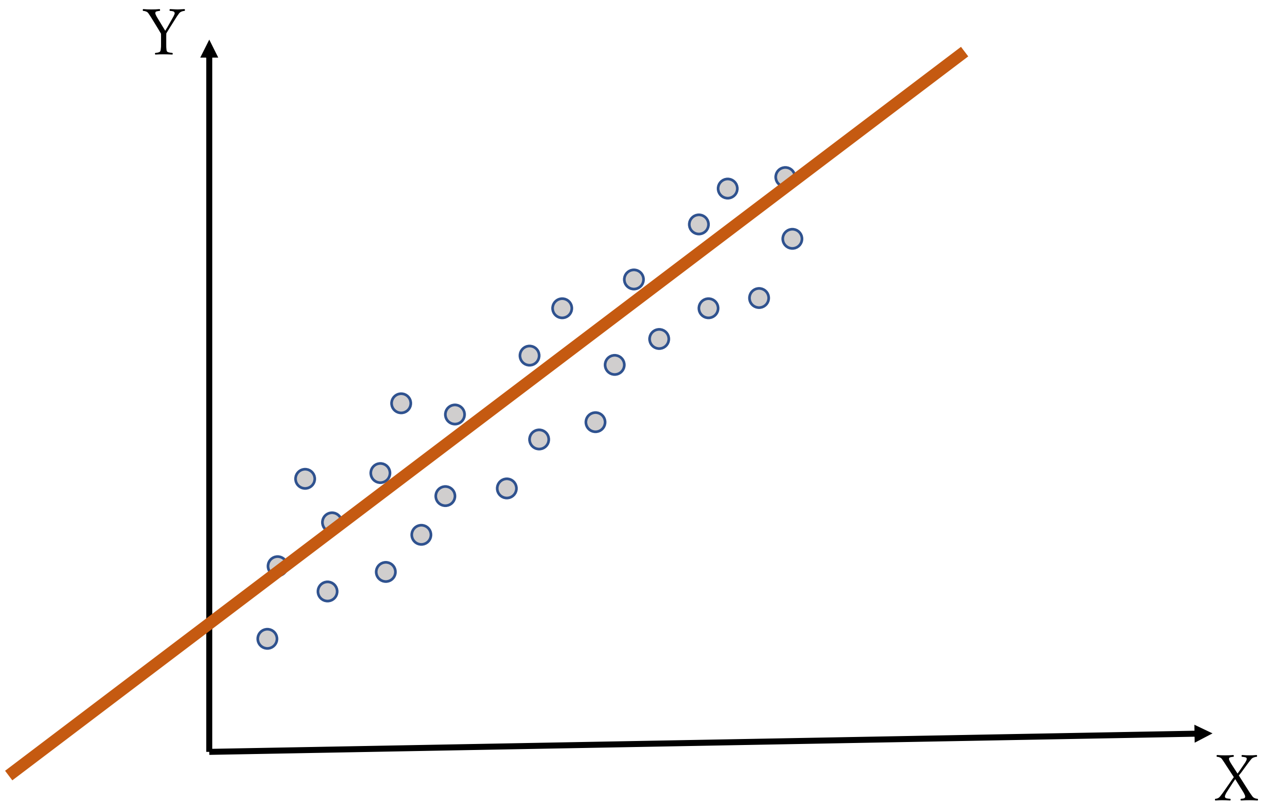
$$Y = \beta_0 + \beta_1 X$$

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X + e$$

$$e = Y - \hat{Y}$$

$$min \sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} + \widehat{\beta_1} X_i\right)^2$$

Is all about solving this equation.
An optimization problem



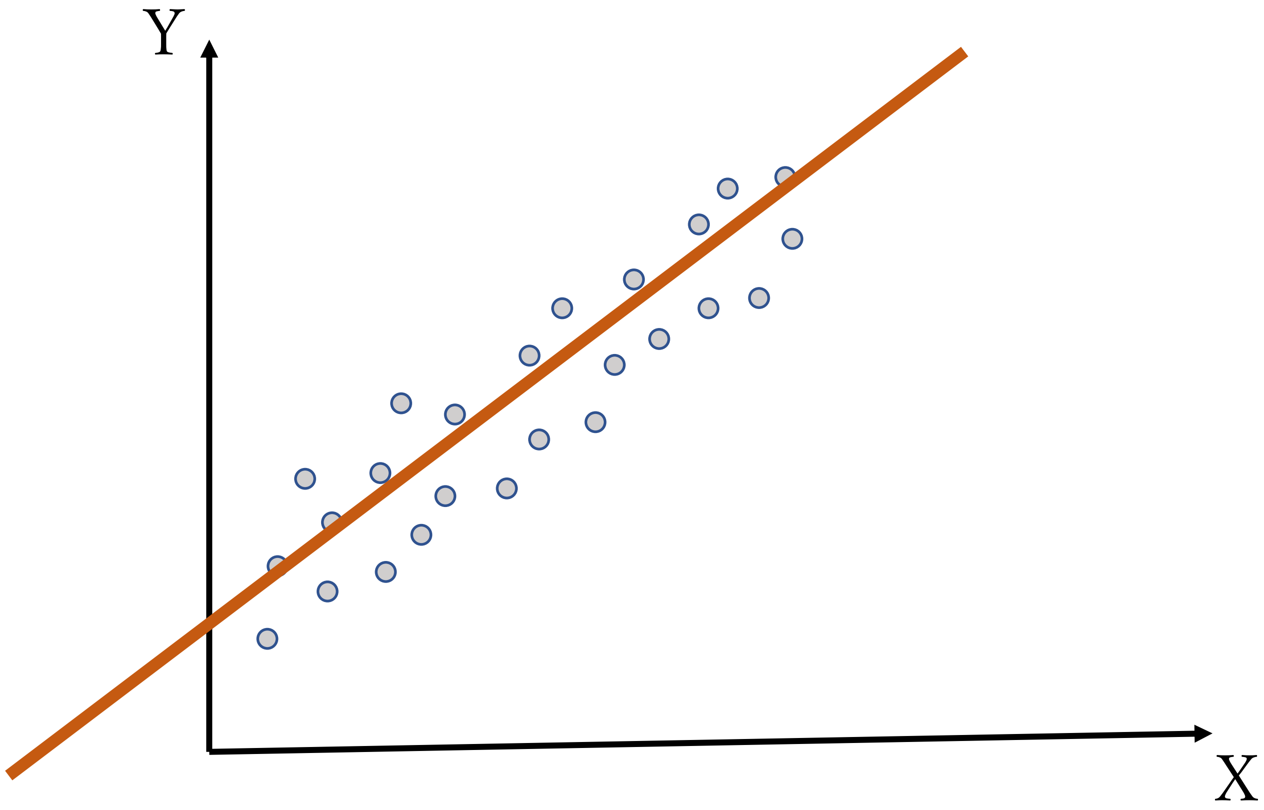
$$Y = \beta_0 + \beta_1 X$$

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X + e$$

$$e = Y - \hat{Y}$$

$$min \sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} + \widehat{\beta_1} X_i\right)^2$$

Is all about solving this equation. An optimization problem



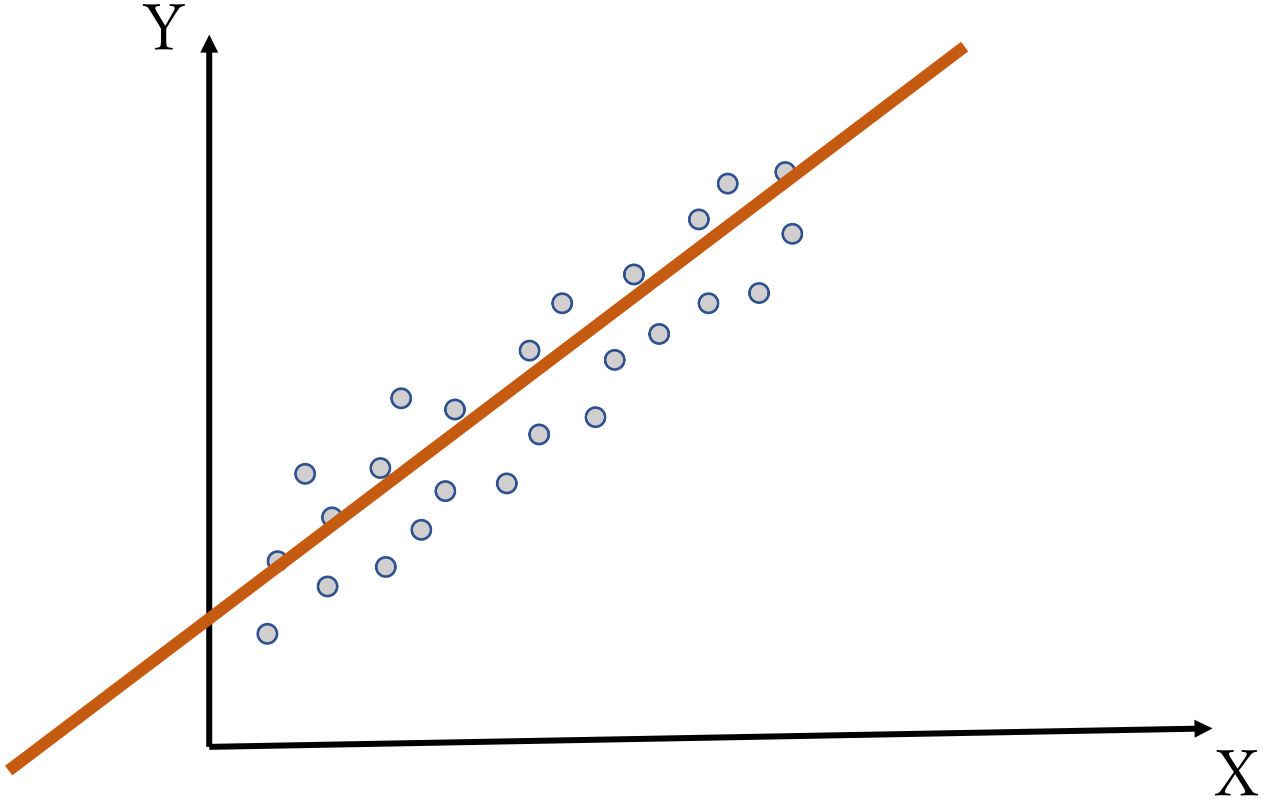
$$S\left(\widehat{\beta_0}, \widehat{\beta_1}\right) = \sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} + \widehat{\beta_1} X_i\right)^2$$

$$\left.\begin{aligned} \frac{\partial S\left(\widehat{\beta_0}, \widehat{\beta_1}\right)}{\partial \widehat{\beta_1}} &= 0 \\ \frac{\partial S\left(\widehat{\beta_0}, \widehat{\beta_1}\right)}{\partial \widehat{\beta_0}} &= 0 \end{aligned}\right\} \text{Normal equations}$$

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

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X + e$$

$$e = Y - \hat{Y}$$



$$min \sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} + \widehat{\beta_1} X_i\right)^2$$

Is all about solving this equation. An optimization problem

$$S(\widehat{\beta_0}, \widehat{\beta_1}) = \sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} + \widehat{\beta_1} X_i\right)^2$$

$$\left.\begin{aligned} \frac{\partial S(\widehat{\beta_0}, \widehat{\beta_1})}{\partial \widehat{\beta_1}} &= 0 \\ \frac{\partial S(\widehat{\beta_0}, \widehat{\beta_1})}{\partial \widehat{\beta_0}} &= 0 \end{aligned}\right\} \text{Normal equations}$$

$$\boxed{\begin{array}{c}\text{This is known as} \\ \text{“Ordinary Least Squares”} \\ \text{OLS}\end{array}}$$

OLS

Given that:

$$\bar{Y} = \hat{\beta}_o + \hat{\beta}_1 \bar{X} \qquad \sigma_{xy} = \sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y}) \qquad \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

We can probe that:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{X})^2} = \frac{\sigma_{xy}}{{\sigma_x}^2} = \rho_{xy} \frac{\sigma_y}{\sigma_x}}$$

$$\boxed{\hat{\beta}_o = \bar{Y} - \hat{\beta}_1 \bar{X}}$$

$$var(\hat{\beta}_1) \qquad var(\hat{\beta}_o)$$

General linear model (or multiple linear regression)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_m X_m$$

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_m x_{ij}$$

General linear model (or multiple linear regression)

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_m x_{ij}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{21} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

General linear model (or multiple linear regression)

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_m x_{ij}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{21} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \hat{\beta}_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{21} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

General linear model (or multiple linear regression)

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_m x_{ij}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{21} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \longrightarrow \boldsymbol{Y} = \boldsymbol{X\beta}$$

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \hat{\beta}_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{21} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \longrightarrow \widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{e}$$

Back to OLS

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \longrightarrow \boldsymbol{SSR} = \mathbf{e'e} \longrightarrow \frac{\delta \boldsymbol{SSR}}{\partial \widehat{\boldsymbol{\beta}}} = 0$$

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{X'X}^{-1}}\boldsymbol{X'Y}$$

Back to OLS

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \longrightarrow \boldsymbol{SSR} = \mathbf{e'e} \longrightarrow \frac{\delta \boldsymbol{SSR}}{\partial \widehat{\boldsymbol{\beta}}} = 0$$

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{X'X}^{-1}}\boldsymbol{X'Y}$$

$$var(\boldsymbol{\beta}) = E\left[(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'\right]$$

$$var(\boldsymbol{\beta}) = \boldsymbol{\sigma_e}^2(\boldsymbol{X'X})^{-1}$$

$$\boldsymbol{\sigma_e}^2 : unknown$$

$$\hat{\sigma}_e{}^2 = S^2 = \frac{\boldsymbol{SSR}}{n - m}$$

Back to OLS

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{X'X^{-1}}}\boldsymbol{X'Y}$$

<u>Consider that:</u>

- **X** is exogenous
- Given **X**, the conditional expected value of **Y** is Xβ and the expected value of **e** is 0.
- The errors are independent and identically distributed
- There is an inverse matrix for **X'X**

BLUE: Best linear unbiased estimator
Gauss-Markov theorem

- Errors are normally distributed

Makes it possible to perform hypothesis tests

A note on estimation

Ordinary least squares → Two least stage squares
2LSS
Weighted least squares WLS
Feasible generalized least squares FGLS

<u>**Maximum likelihood estimation**</u> → A statistical method used to estimate the parameters of a probability distribution by maximizing the likelihood function, which is a function that measures how likely it is to observe the data given the parameters of the distribution. The basic idea behind MLE is to find the parameter values that make the observed data the most probable

The assumptions

Linearity

Independence

Homoscedasticity

Normality

The assumptions

Linearity → In the parameters

Independence → Residuals

Homoscedasticity → Residuals

Normality → Residuals

The assumptions

Linearity	→	In the parameters		
Independence	→	Residuals	→	Durbin Watson
Homoscedasticity	→	Residuals	→	Q-Q plot, Jarque-Berra, Breusch-Pagan, Koenker, White
Normality	→	Residuals	→	Kolmogorov-Smirnov, Shapiro-Wilk

You should also check for multicollinearity → Variation Inflation Factor VIF

Linearity → In the parameters

Independence → Residuals → Durbin Watson

Homoscedasticity → Residuals → Q-Q plot, Jarque-Berra, Breusch-Pagan, Koenker, White

Normality → Residuals → Kolmogorov-Smirnov, Shapiro-Wilk

So far, what does the output look like?

Variables	Parameter	Standard Error	t value	p value	
Intercept	---	---	---	---	*
X1	---	---	---	---	
X2	---	---	---	---	*
X2	---	---	---	---	
X4	---	---	---	---	***
X5	---	---	---	---	**

R-squared and adjust R-squared are also presented
Significance of parameter: t-statistic
Significance of parameter (with reference to a value): Wald test
Overall significance of the regression (all parameters = 0): F-test

A note on violation of the assumptions

- Nonlinearity: try transformations, check for outliers or influential points, and other types of regression

- Non-independence: transformations, add/remove variables, spatial regression models that control for spatial autocorrelation

- Non-normality: transformations

- Heteroscedasticity: transformations, robust standard errors, Weighted OLS, Quantile regression, Geographically Weighted Regression GWR

- Model specification: knowledge domain, endogeneity (correlation of a variable with the error term), omitted variable bias (variable related to ta independent variable and the outcome)

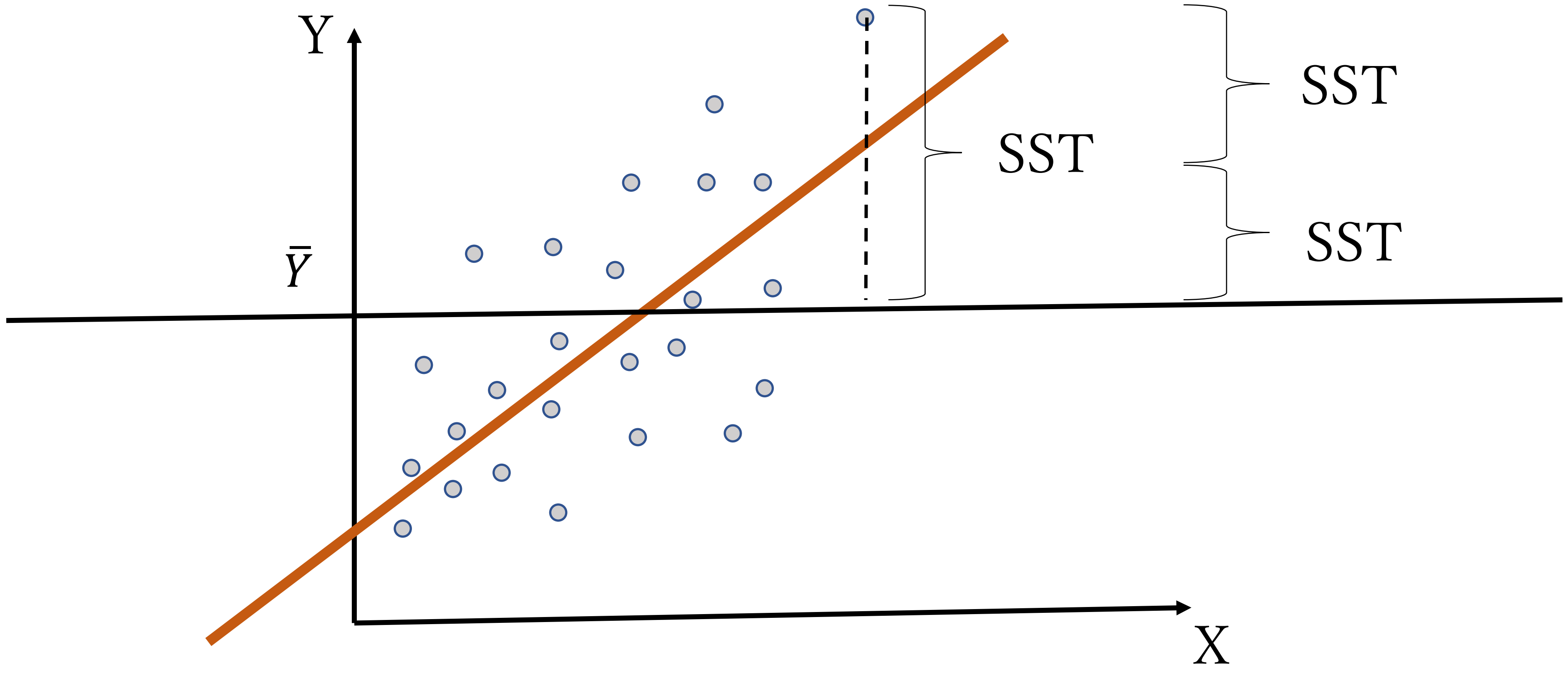
R-squared (R2) and adjusted R-squared

- Measures of how well the regression model fits the observed data
- They provide information about the proportion of variance explained by the model and help evaluate its potential predictive power

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad R^2_{adj} = 1 - \left(\frac{n - 1}{n - p - 1}\right)(1 - R^2)$$

SSR (Sum of Squares Regression) is the sum of squared differences between the predicted values and the mean of the dependent variable.
SSE (Sum of Squares Error) is the sum of squared residuals, which are the differences between the observed values and the predicted values.
SST (Sum of Squares Total) is the total sum of squares, which is the sum of squared differences between the observed values and the mean of the dependent variable.
n is the sample size (number of observations).
p is the number of predictor variables in the mode.

R-squared (R2) and adjusted R-squared



Variance Inflation Factor (VIF)

- Assess multicollinearity
- The VIF measures how much the variance of the estimated regression coefficient is inflated due to multicollinearity. It quantifies the extent to which the variance of an estimated regression coefficient is increased compared to the situation when there is no multicollinearity

$$VIF_j = \frac{1}{1 - R_j^2}$$

- VIF = 1: No multicollinearity.
- VIF > 1 and < 5: Moderate multicollinearity.
- VIF > 5: High multicollinearity.

VIF_j: $Variance\ Inflation\ Factor\ for\ predictor\ variable\ X_j$

R_j^2: $coefficient\ of\ determination\ (R - squared)\ from\ regressing\ X_j\ \ on\ all\ the\ other\ predictor$

Likelihood and deviance

Saturated model: model where each point hast its own parameters

We can have log likelihoods for:

- a model with no explanatory variables, only the intercept (null model)
- our proposed model
- the saturated mode

$$Deviance_{null} = 2\big(LL(Saturated\ model) - LL(Null\ model)\big)$$

$$Deviance_{residual} = 2\big(LL(Saturated\ model) - LL(Proposed\ model)\big)$$

A good model should have low residual deviance relative to the null deviance

Likelihood ratio test LRT

Mechanism to test if the proposed model provides a significant improvement over the null

The proposed and the null are nested

$$D: Likelihood\ ratio\ test\ statistic$$

$$D = -2ln\left(\frac{LL_{Null\ model}}{LL_{Proposed\ model}}\right) = -\left(ln(LL_{Null\ model})\ - ln(LL_{Proposed\ model})\right)$$

$$D\ =\ Deviance_{null}\ - Deviance_{residual}$$

The likelihood ratio test is assumed to follow a chi-squared distribution. The degrees of freedom are the number of estimated parameters in the proposed model. The null is that the proposed model and the null model are equal. We want to reject the null (p value <0.05)

Akaike's Information Criterion AIC

Mechanism to compare two models even if they are not nested

$$AIC \ = 2p - 2LL$$

p: $number\ of\ parameters\ in\ the\ model$

LL: $loglikelihood$

We select the model with the lowest AIC value

Bayesian Information Criterion AIC

p: $number\ of\ parameters\ in\ the\ model$

$$BIC \ = p * ln(n) - 2LL$$

LL: $loglikelihood$

n : $sample\ size$

We select the model with the lowest BIC value

Categorical variables - interactions

Any idea?

Why is heteroskedasticity an issue?

Homoscedasticity: residuals have mean zero and equal variance at any location of X

- Inefficient parameter estimates: the coefficients may be more influenced by observations with larger variances, leading to inaccurate and inefficient estimates

- Wrong standard errors (biased): the standard errors may be underestimated or overestimated

- Unreliable hypothesis testing: Inaccurate standard errors can lead to incorrect inferences about the statistical significance of the coefficients and affect hypothesis testing and confidence intervals. When the assumption of homoscedasticity is violated, the standard t-tests and p-values may be unreliable, leading to incorrect conclusions about the statistical significance of the relationships between the predictor variables and the dependent variable.

- Inefficient prediction

Thank you

Orlando Sabogal-Cardona
PhD researcher
University College London UCL