

Linear Regression

A scatter plot with a diagonal regression line and the word 'Supervised' in blue. The plot shows a positive correlation between two variables, with a line of best fit drawn through the data points. The word 'Supervised' is written in blue text to the right of the plot.

Supervised

Hisham Ihshaish
Birzeit University
May 2026

Agenda

➡ Motivation

➡ Linear Regression Models

⦿ Estimation

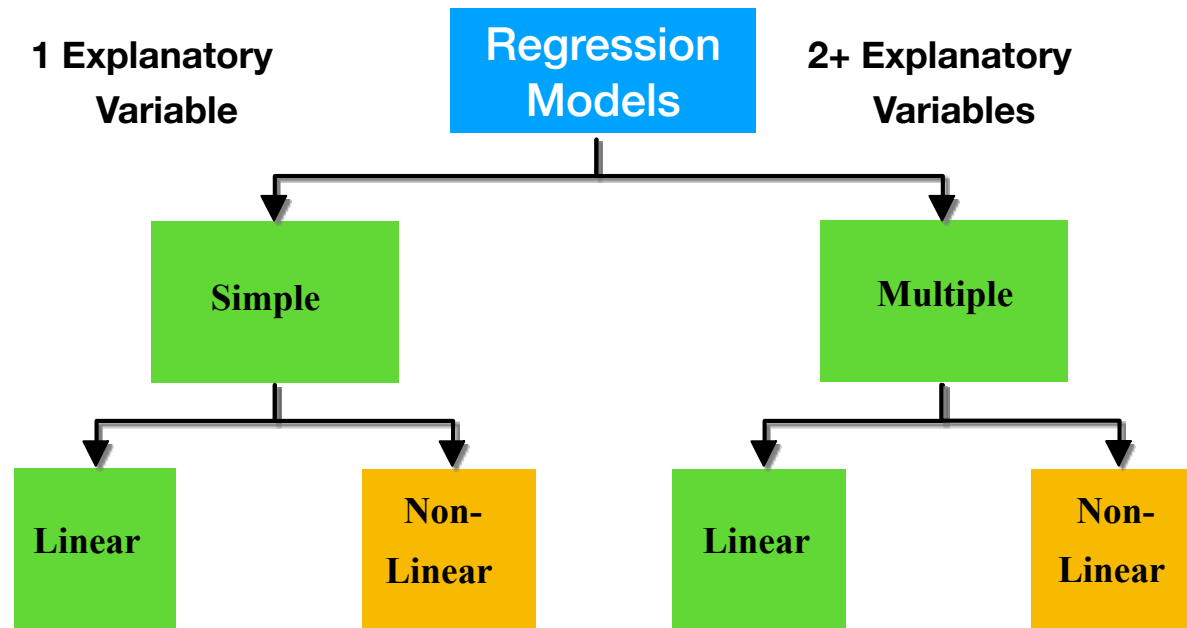
⦿ Evaluation

➡ Tutorial (1) — you will be provided with further exercise later on.

Note: if you see many equations and dislike them, you're absolutely OK (these are just an extra).

I'll show you what we need to minimally understand by the end of this lecture!

Regression Models



Motivation

- ▶ Analyse the *specific* relationships between the independent variables and the dependent one.
- ▶ Predict the value of a *dependent* variable (Y_i) from the value of *independent* variable ($X1_i, X2_i, \dots, Xn_i$)

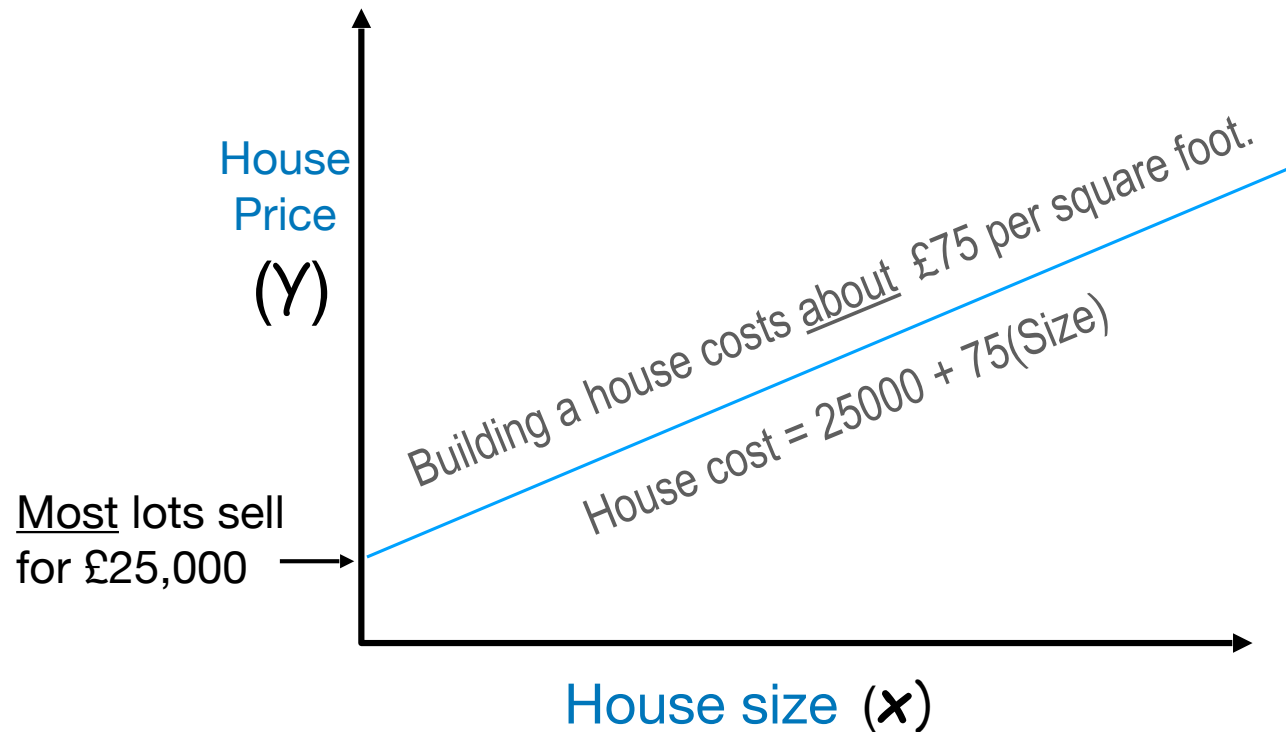
Motivation

- ▶ Predict the value of a dependent variable (Y_i) from the value of independent variable ($X1_i, X2_i, \dots, Xn_i$)
- ▶ Analyse the specific relationships between the independent variables and the dependent one

simple LR

Multiple LR

Relationship between 2 variables

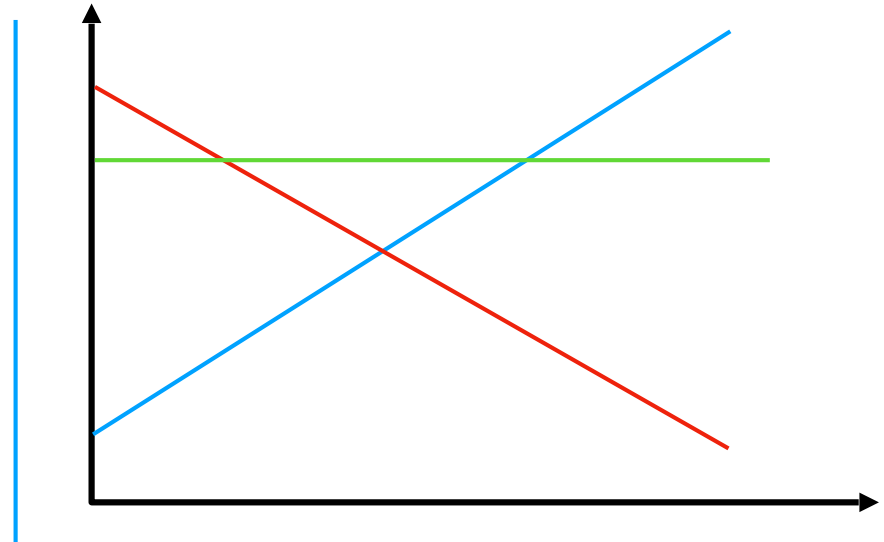


! Extrapolation vs Interpolation

General Model

- $\beta_1 > 0 \Rightarrow$ Positive Association
- $\beta_1 < 0 \Rightarrow$ Negative Association
- $\beta_1 = 0 \Rightarrow$ No Association

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

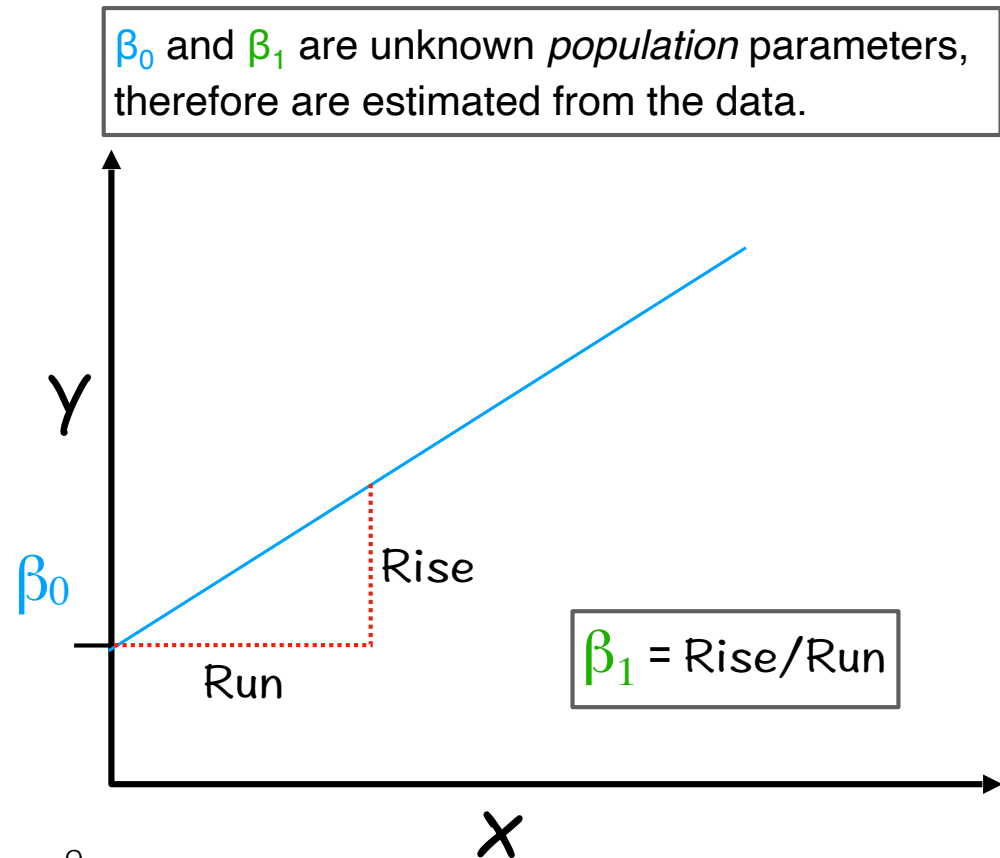


- $\beta_0 \equiv$ Mean response when $x=0$ (y -intercept)
- $\beta_1 \equiv$ Change in mean response when x increases by 1 unit (slope)
- β_0, β_1 are unknown parameters (like μ)
- $\beta_0 + \beta_1 x \equiv$ Mean response when independent variable takes on the value x

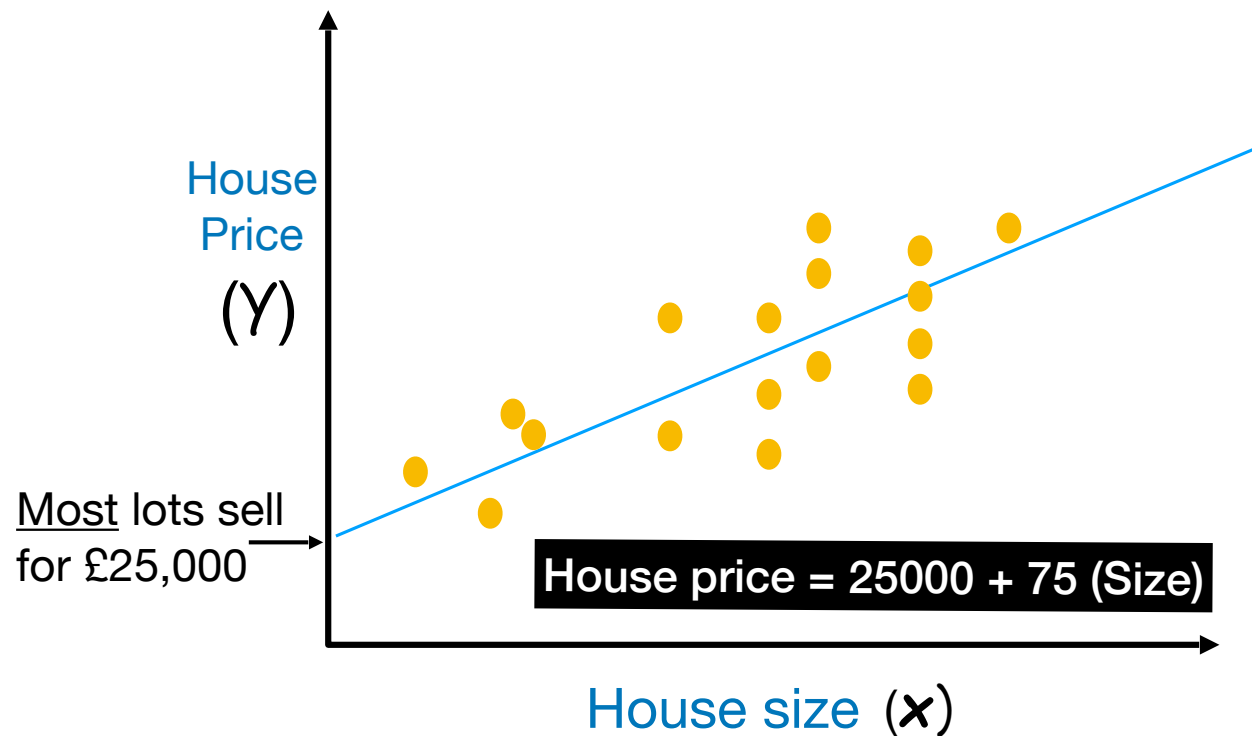
Simplified (simple LR)

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

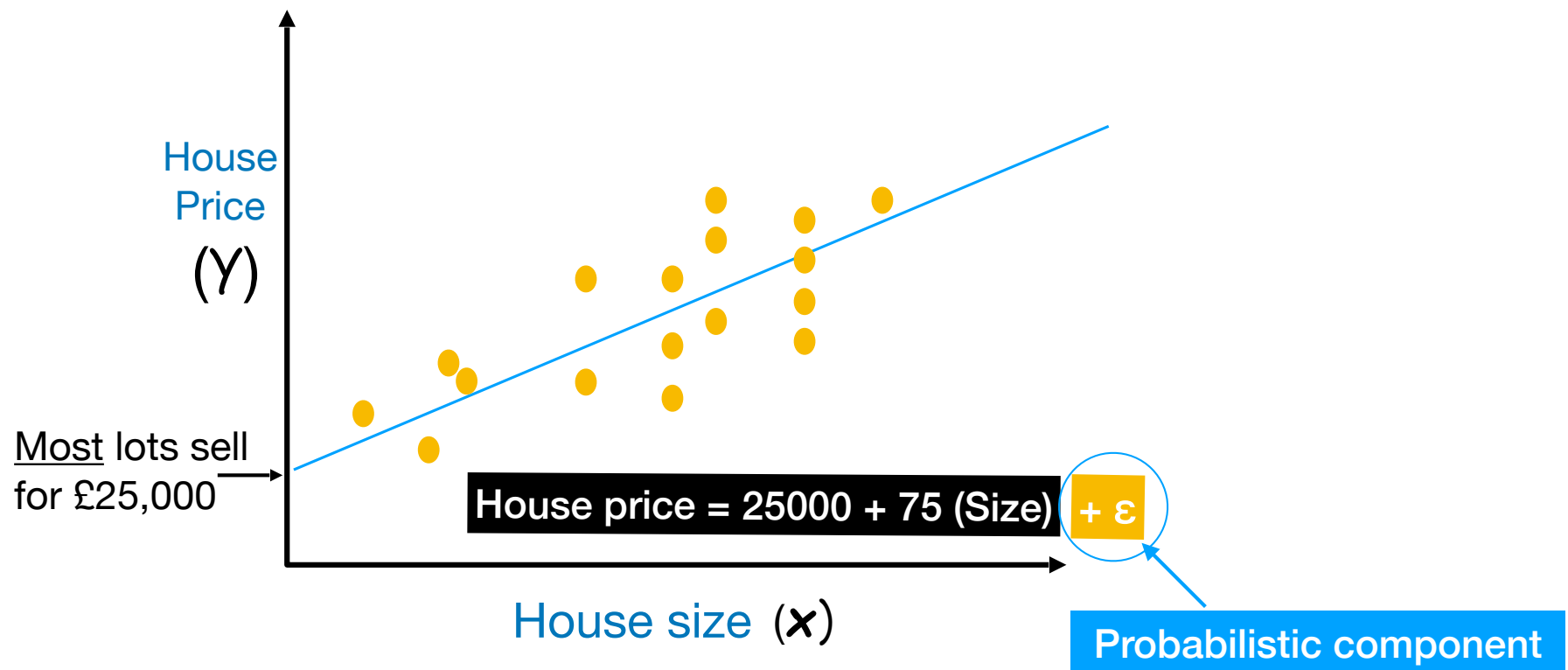
- ▶ y = dependent variable
- ▶ x = independent variable
- ▶ β_0 = y-intercept
- ▶ β_1 = slope
- ▶ ε = error variable



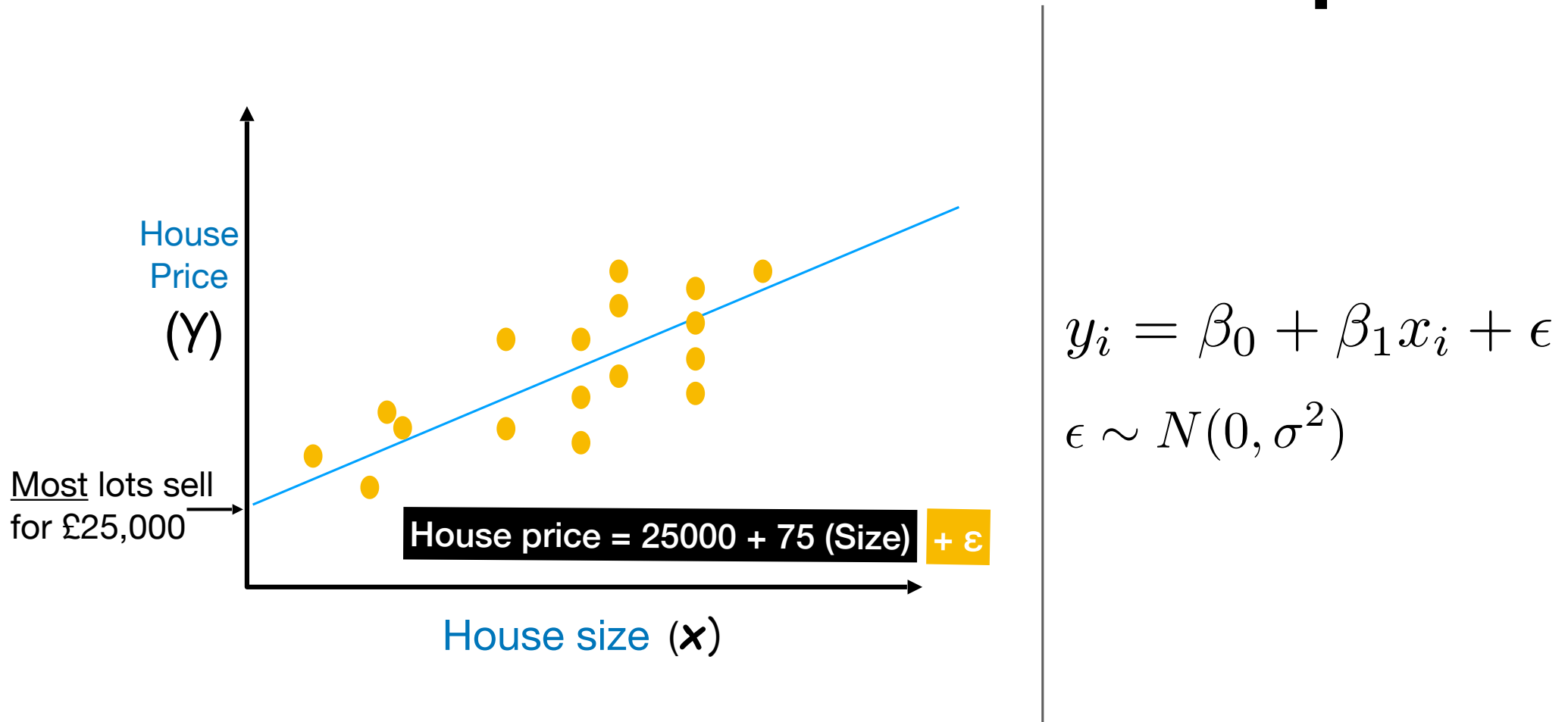
Deterministic and Probabilistic relationship



Deterministic and Probabilistic relationship

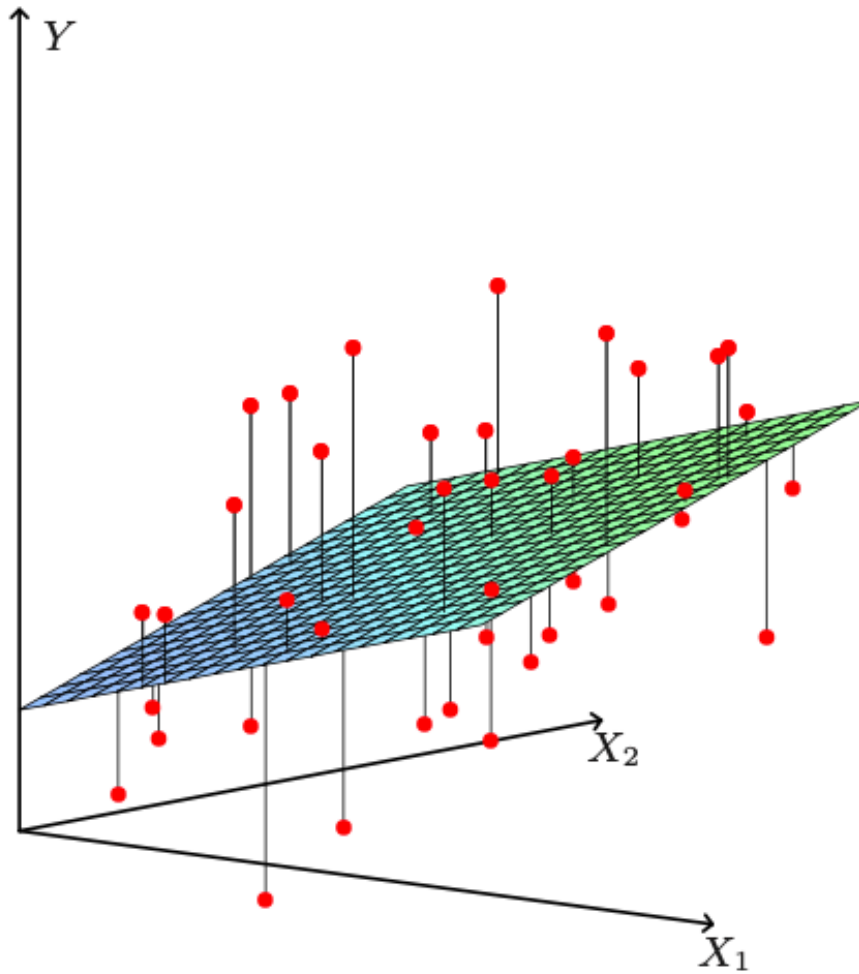


Deterministic and Probabilistic relationship



Note that the intercept β_0 , the slope β_1 , and the noise variance σ^2 are all treated as fixed (i.e., deterministic) but unknown quantities.

Motivation



$$h(x) = \beta_0 + \sum_{i=1}^P X_i \beta_i$$

Linear least squares fitting with $X \in \mathbb{R}^p$. We seek the linear function of X that minimises the sum of squared residuals from Y .

Assumptions

(1) The regression model is linear in parameters (Weak exogeneity)

⦿ $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

⦿ $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$

(2) Mean of residuals is zero & Homoscedasticity of constant variance

⦿ $\epsilon \sim N(0, \sigma^2)$

(3) Feature vectors and residuals are uncorrelated.

(4) No perfect multicollinearity

⦿ Variance Inflation Factor (VIF) $VIF_i = \frac{1}{(1 - R_i^2)}$

`from statsmodels.stats.outliers_influence import variance_inflation_factor`

Agenda

➔ Motivation

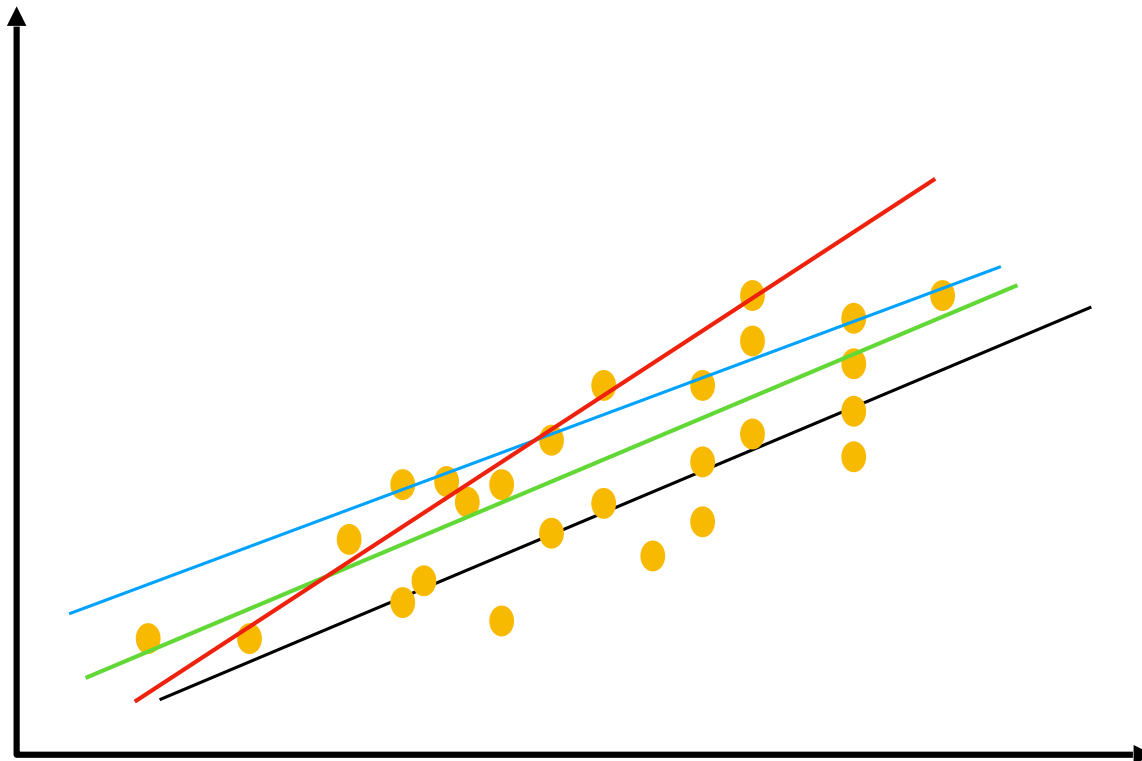
➔ Linear Regression Models

● Estimation

● Evaluation

➔ Tutorial (1) — you will be provided with further exercise later on.

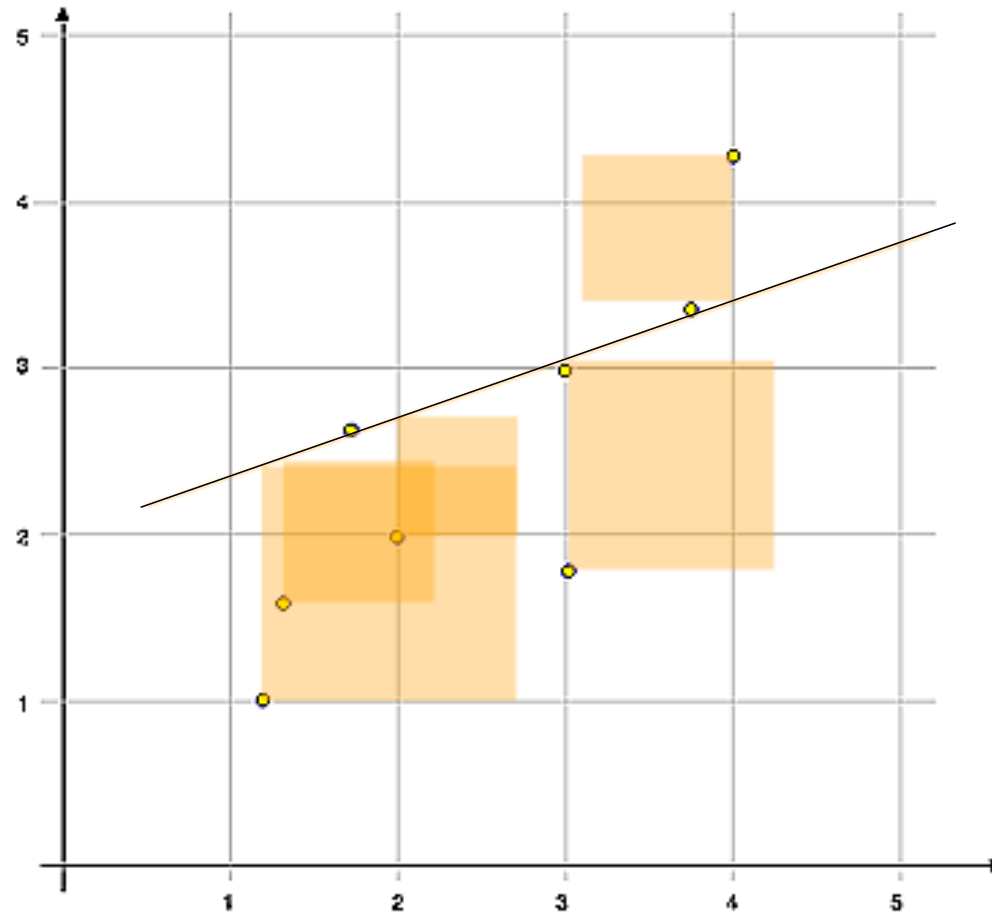
Which Line (model) then?



β_0, β_1

1) OLS {ordinary least squares} //

LSE **Least Square Error**



1) MLE {Maximum Likelihood Estimate}

The Assumptions again!

1. The distribution of X is arbitrary (and perhaps X is even non-random).
2. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some constants (“coefficients”, “parameters”) β_0 and β_1 , and some random noise variable ϵ .
3. $\epsilon \sim N(0, \sigma^2)$, and is independent of X .
4. ϵ is independent across observations.

$$P(y_i | X = x_i; \beta_0, \beta_1, \sigma^2)$$

Given any data set $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, we can now write down the probability density, under the model, of seeing that data:

$$P(y_i | X = x_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$$

1) MLE {Maximum Likelihood Estimate}

$$P(y_i | X = x_i; \beta_0, \beta_1, \sigma^2)$$

$$P(y_i | X = x_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$$

$$\arg \max_{\vec{\beta}} \prod_{i=1}^N P(y_i | x_i, \vec{\beta})$$

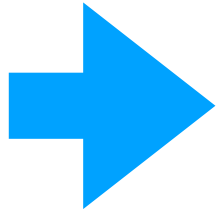
$$= \log \prod_{i=1}^n p(y_i | x_i; b_0, b_1, s^2) \quad (1)$$

$$\arg \max_{\vec{\beta}} \sum_{i=1}^N \log P(y_i | x_i, \vec{\beta})$$

$$= \sum_{i=1}^n \log p(y_i | x_i; b_0, b_1, s^2) \quad (2)$$

$$= -\frac{n}{2} \log 2\pi - n \log s - \frac{1}{2s^2} \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2 \quad (3)$$

... "under certain assumptions any learning algorithm that minimizes the squared error between the output hypothesis predictions and the training data will output a maximum likelihood hypothesis." – T. Mitchell *Machine Learning* [1st Ed.], 1997.



Simple linear regression

Solving for the fit: least-squares regression

Assuming that this is actually how the data $(x_1, y_1), \dots, (x_n, y_n)$ we observe are generated, then it turns out that we can find the line for which the probability of the data is highest by solving the following optimization problem * :

$$\min_{\beta_0, \beta_1} : \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2, \quad (1)$$

where \min_{β_0, β_1} means “minimize over β_0 and β_1 ”. This is known as the **least-squares linear regression problem**. Given a set of points, the solution is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2} \rightarrow \frac{\text{cov}(x, y)}{s_x^2} \quad (2)$$

$$= r \frac{s_y}{s_x}, \quad (3)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad (4)$$

* This is an important point: the assumption of Gaussian noise leads to squared error as our minimization criterion.

➔ Gradient Descent {ordinary least squares}

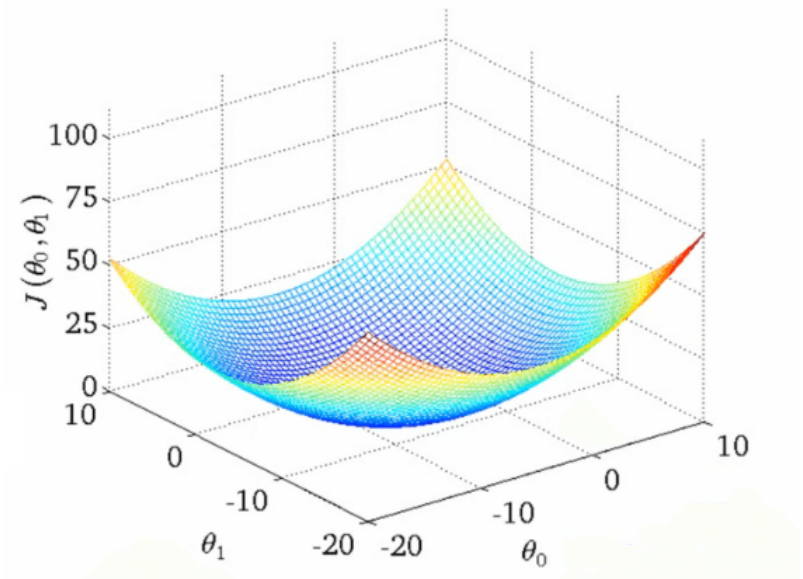
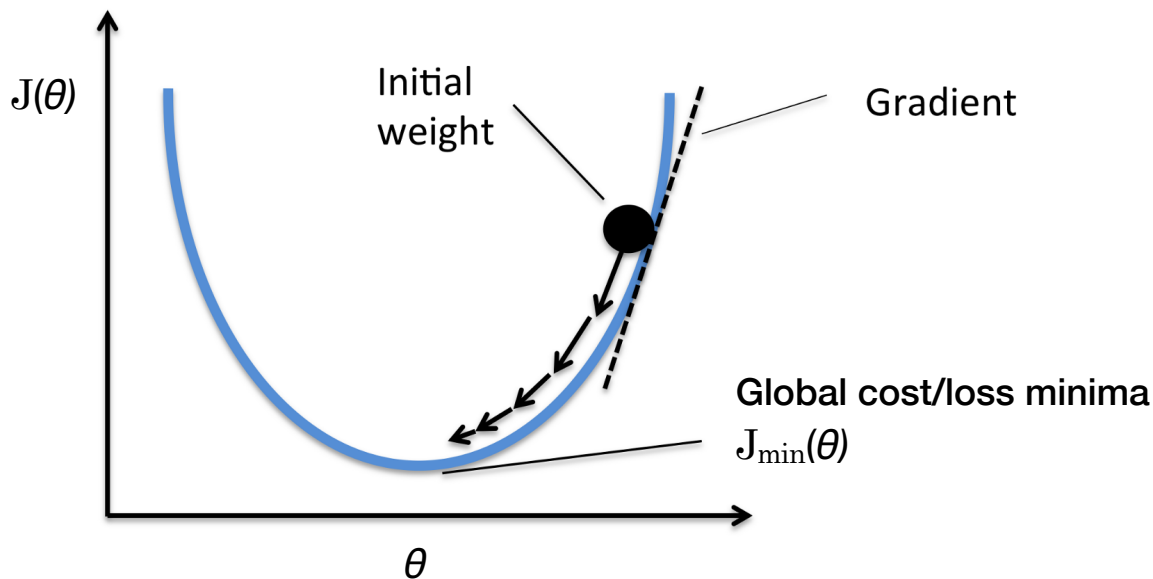
Let θ_0 and θ_1 be your linear model (parameters, previously we called them β_0, β_1)

Gradient descent algorithm

repeat until convergence {
 $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$
(for $j = 1$ and $j = 0$)
}

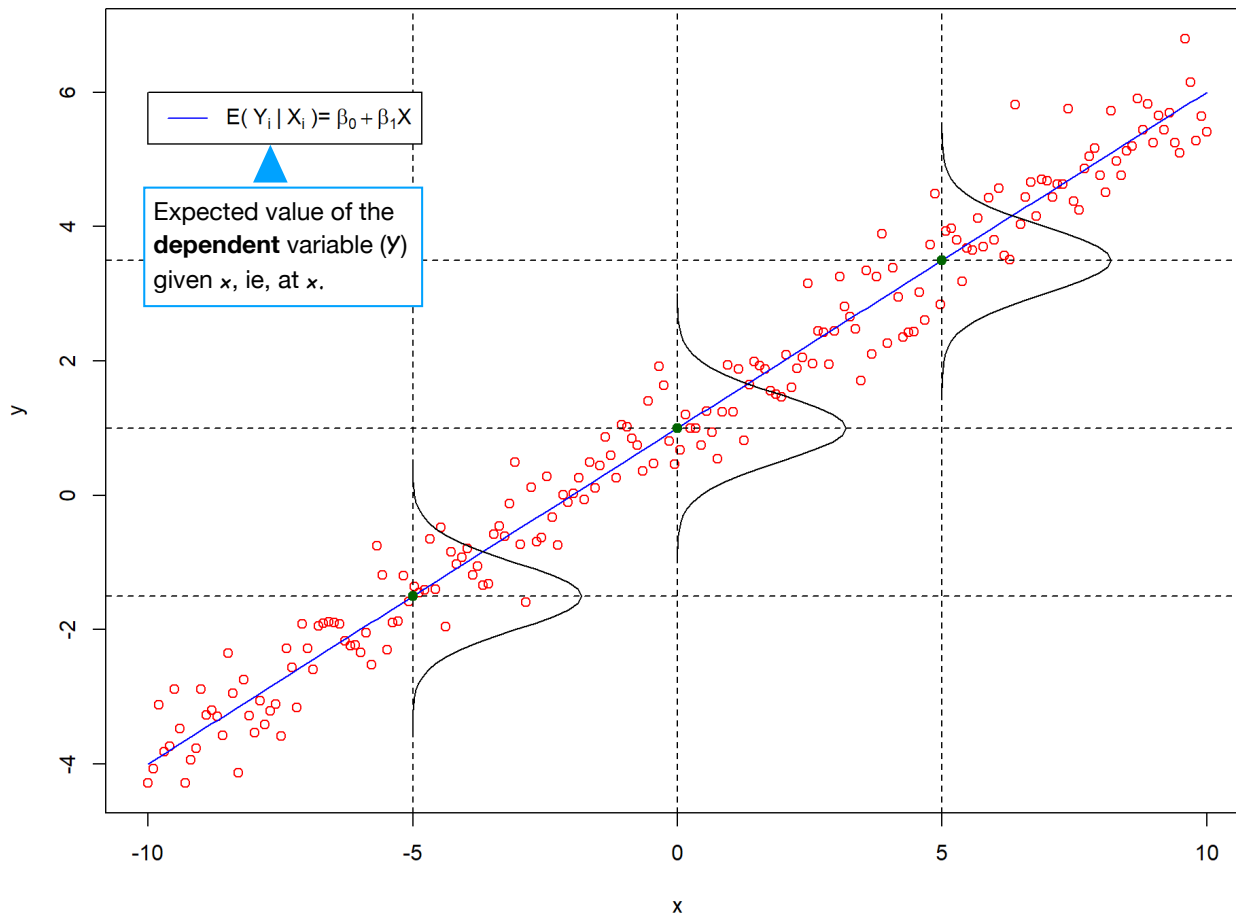
Linear Regression Model

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$



What to takeaway!

Linear Regression with Gaussian Errors



$$E(Y | \mathbf{x}) = \beta_0 + \beta_1 \mathbf{x}$$



The structural model underlying a linear regression analysis is that the **explanatory** (\mathbf{x}) and outcome variables (y) are linearly related such that the population mean of the outcome for any \mathbf{x} value is $\beta_0 + \beta_1 \mathbf{x}$.

What to takeaway!

- ➔ Explanatory (independent) and Response (dependent) Variables are *Numeric*.
- ➔ Relationship between the ***mean*** of the response variable and the values of the explanatory variable assumed to be approximately linear (straight line)

Agenda

➔ Motivation

➔ Linear Regression Models

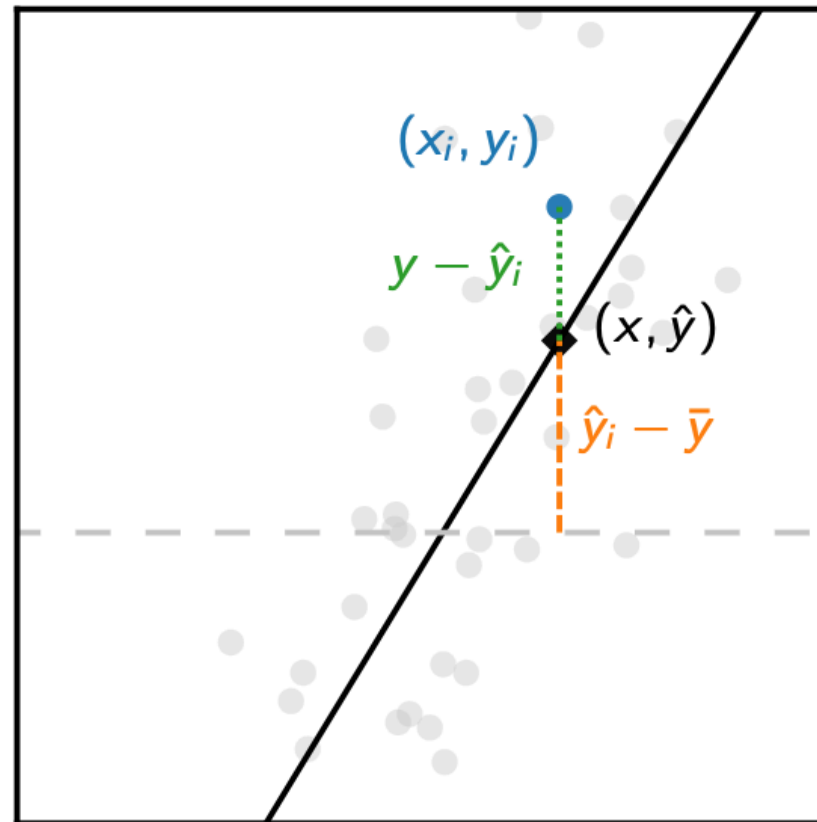
⦿ Estimation

● Evaluation

➔ Tutorial (1) — you will be provided with further exercise later on.

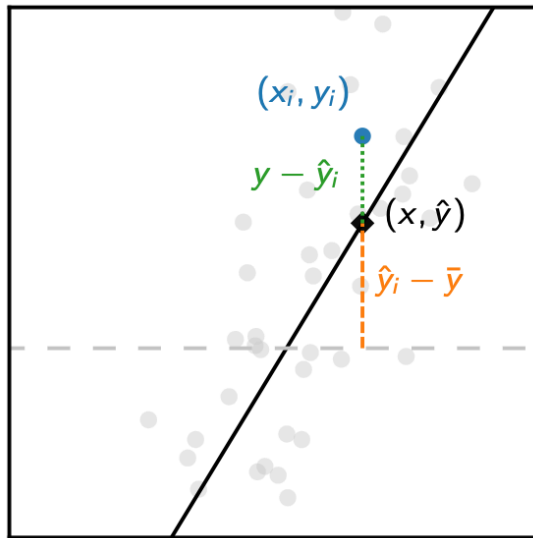
Evaluating LR model

Total LR square error



How much (what %) of the variation in y (**orange** and **green**) is NOT explained by the variation in x (**as in LR model**)? — **green left out** (ie, unexplained by the model).

Coefficient of Determination



$$y_i - \bar{y} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{difference explained by model}} + \underbrace{(y_i - \hat{y}_i)}_{\text{difference not explained by model}}$$

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{SS_{\text{total}} \downarrow SS_{yy}} = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{SS_{\text{model}}} + \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{SS_{\text{error}}}$$

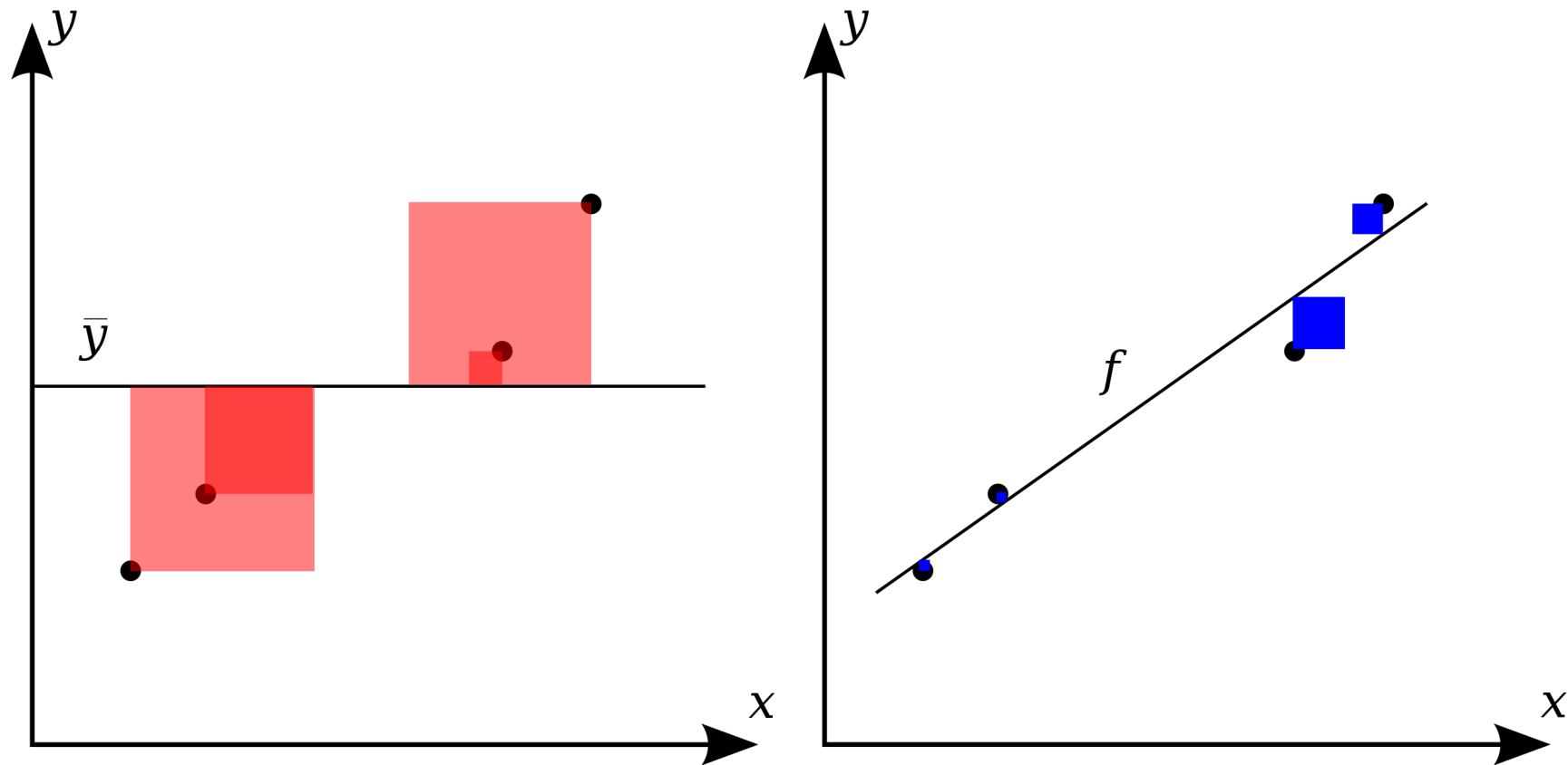
Coefficient of Determination $[0,1]$

the higher, the best

$$\rightarrow R^2 = (R)^2 = 1 - \frac{SSE}{SS_{yy}}$$

Note that $SS_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$ represents the total variation in the response variable, while $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ represents the variation in the observed responses about the fitted equation (after taking into account x). This is why we sometimes say that R^2 is “proportion of the variation in y that is ‘explained’ by x .”

Coefficient of Determination



$$R^2 = (R)^2 = 1 - \frac{SSE}{SS_{yy}}$$

The better the linear regression (on the right) fits the data in comparison to the simple average (on the left graph), the closer the value of R^2 is to 1. The areas of the blue squares represent the squared residuals with respect to the linear regression. The areas of the red squares represent the squared residuals with respect to the average value (mean of observations). — [From Wikipedia on Coefficient of determination](#).

Exercise!

What is the Adjusted R^2 — explore and apply on tasks B2 of the Tutorial.

Residual Plots

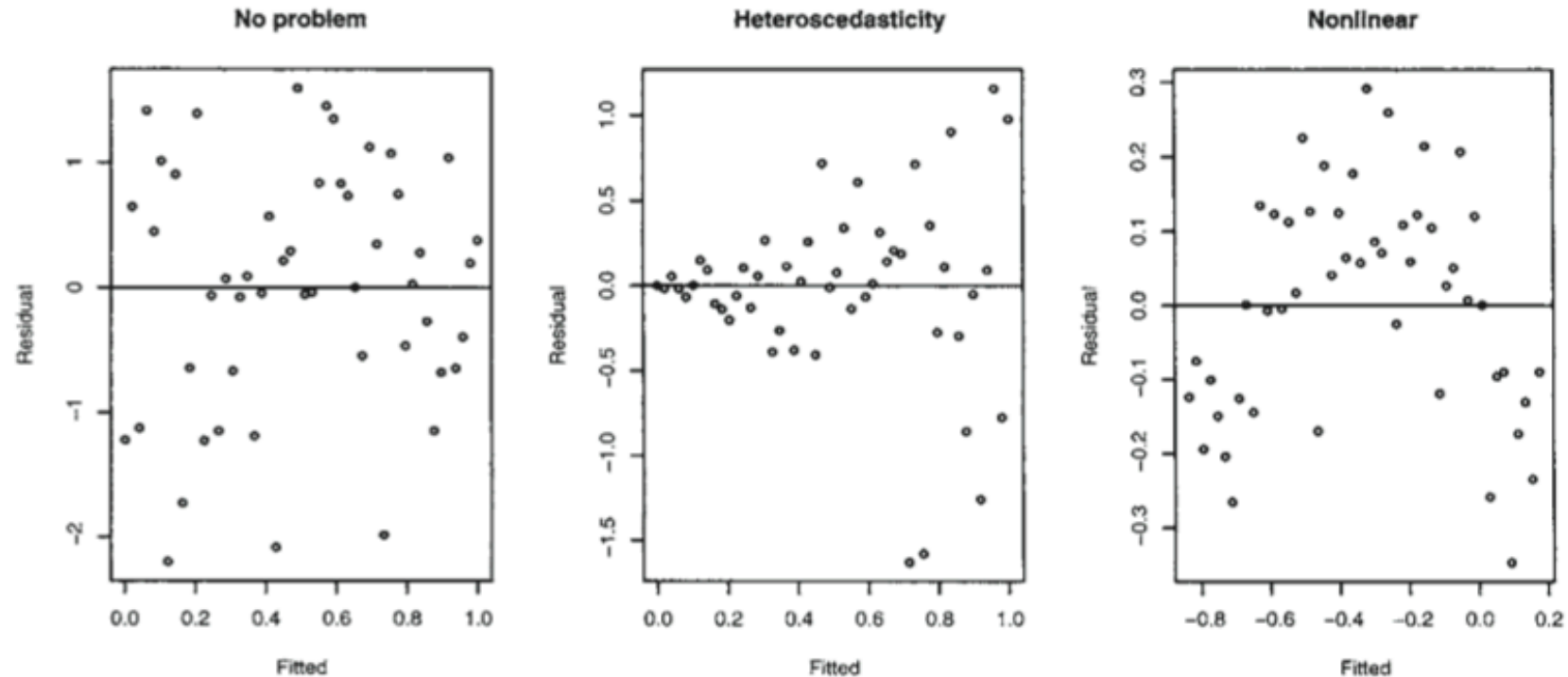
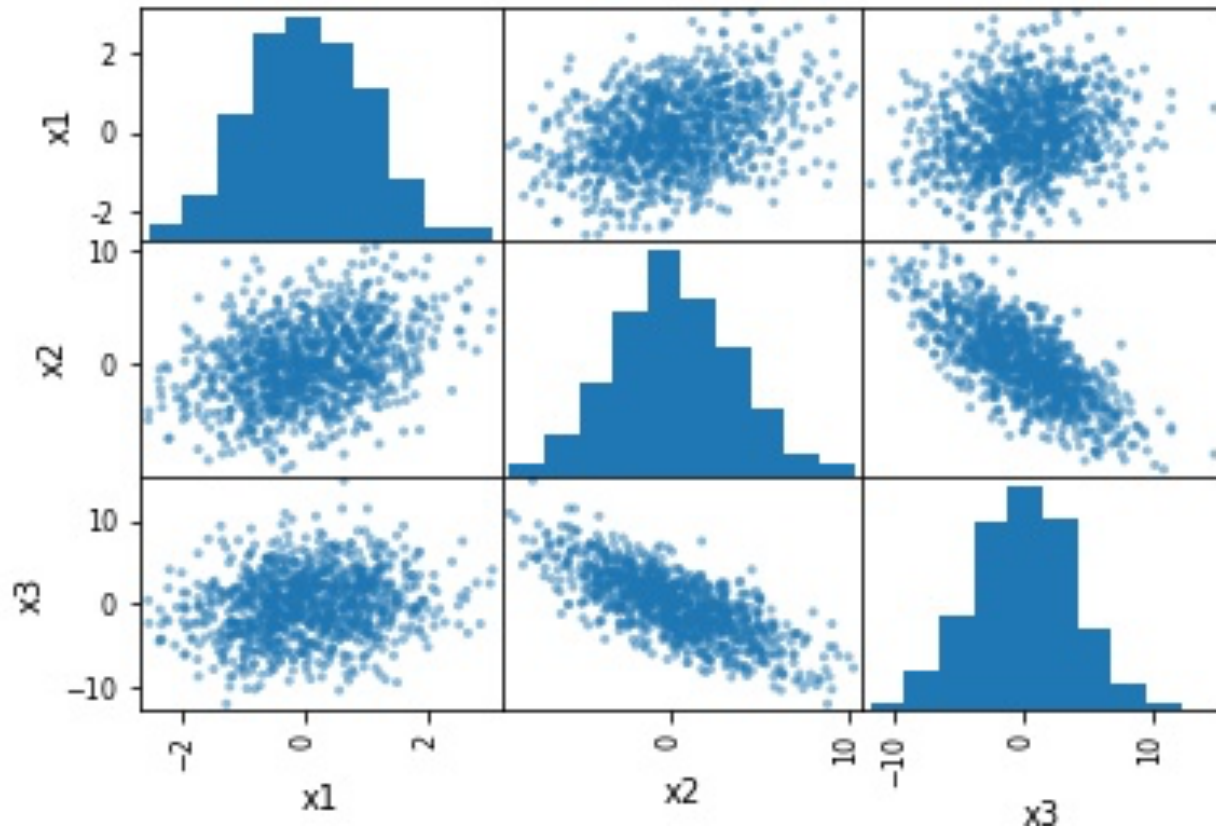


Figure 4.1 *Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.*

Choosing independent variables - Predictors!

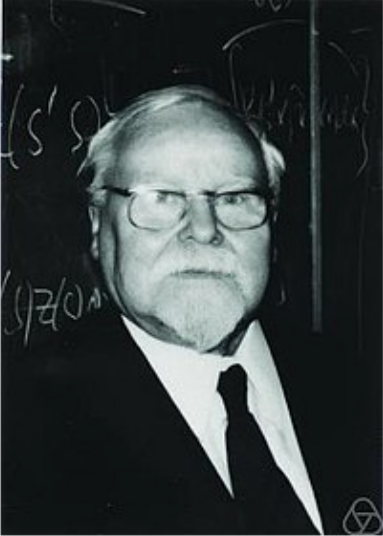


Correlation Matrix

Multicollinearity

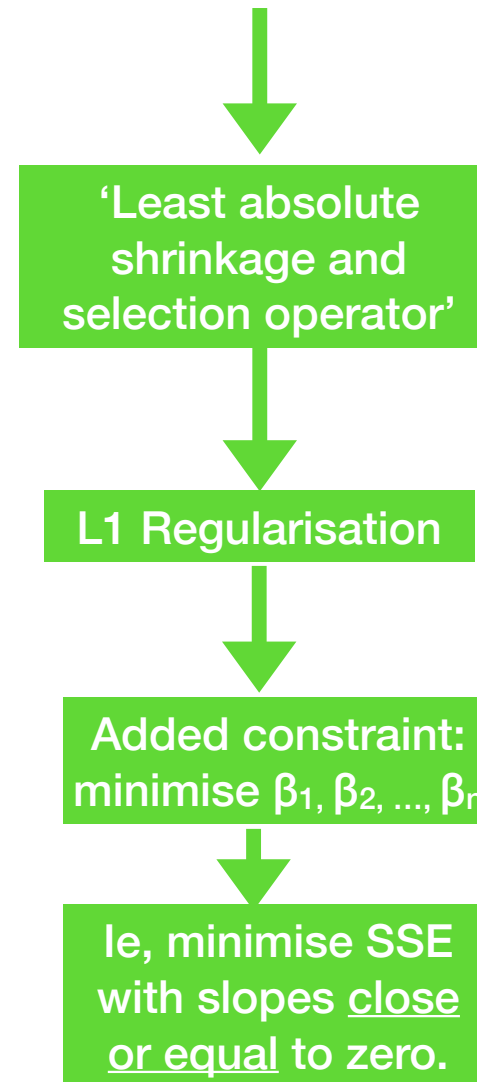
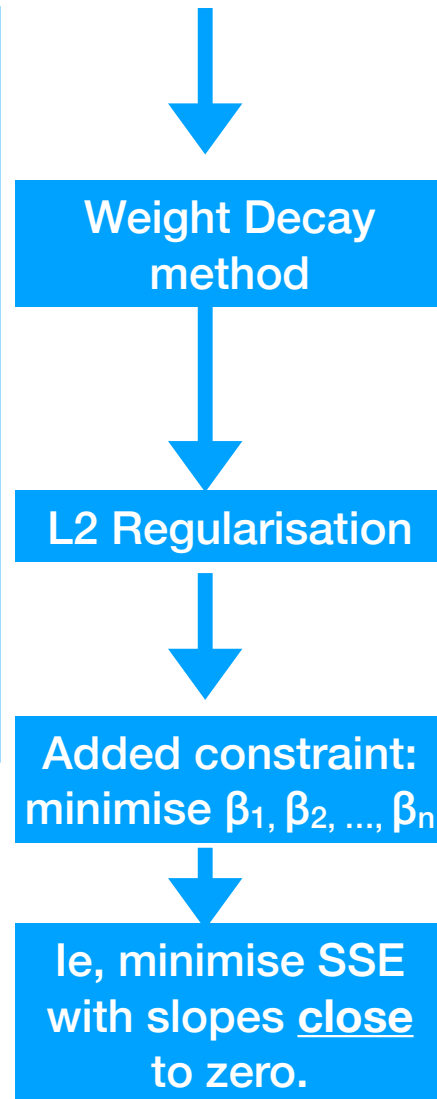
Ridge and Lasso

Andrey Tikhonov




Born 30 October 1906
Gzhatsk, Russian Empire

Died October 7, 1993 (aged 86)
Moscow, Russia



Robert Tibshirani



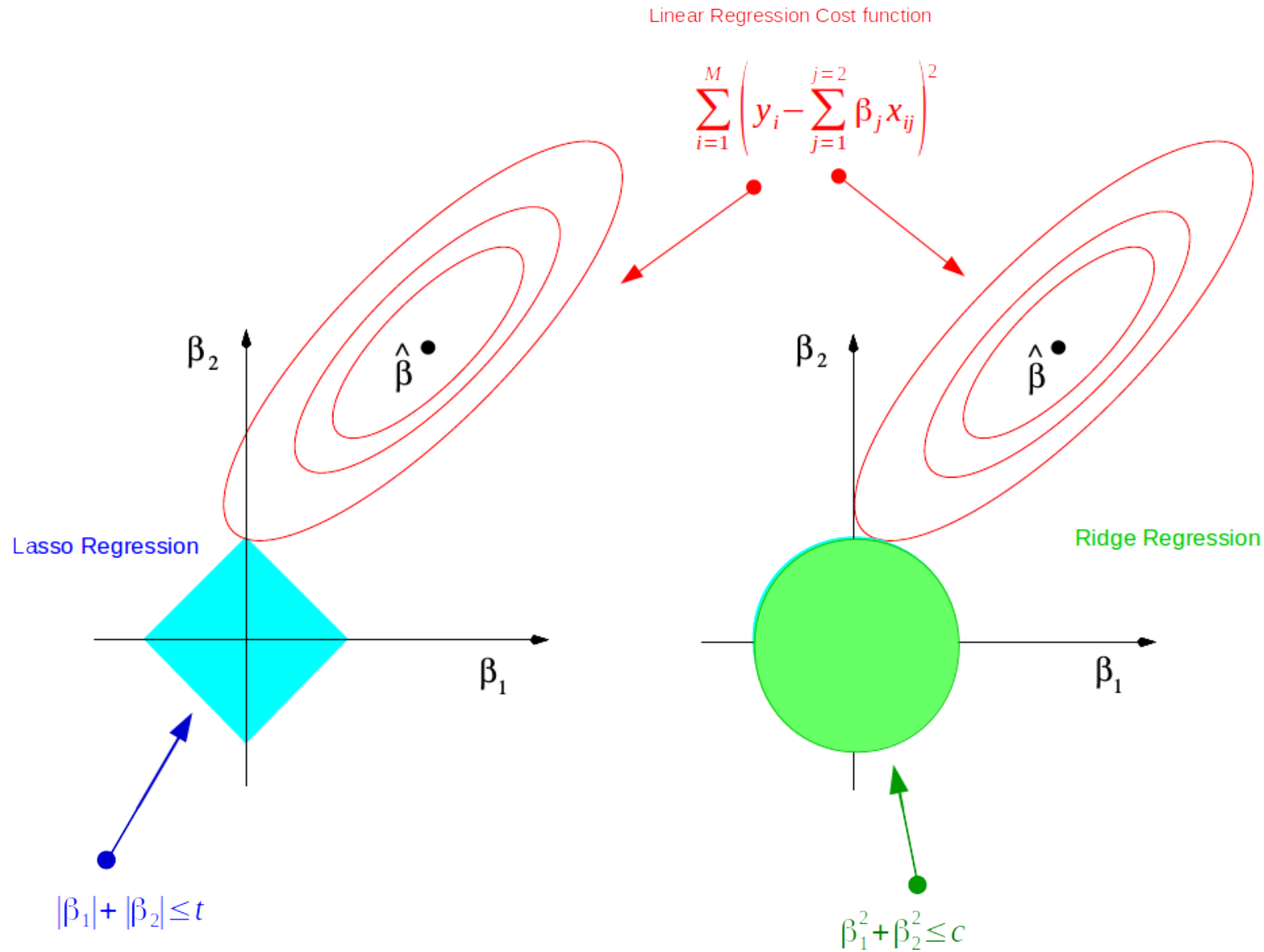
Born July 10, 1956 (age 61)

Nationality Canada, American

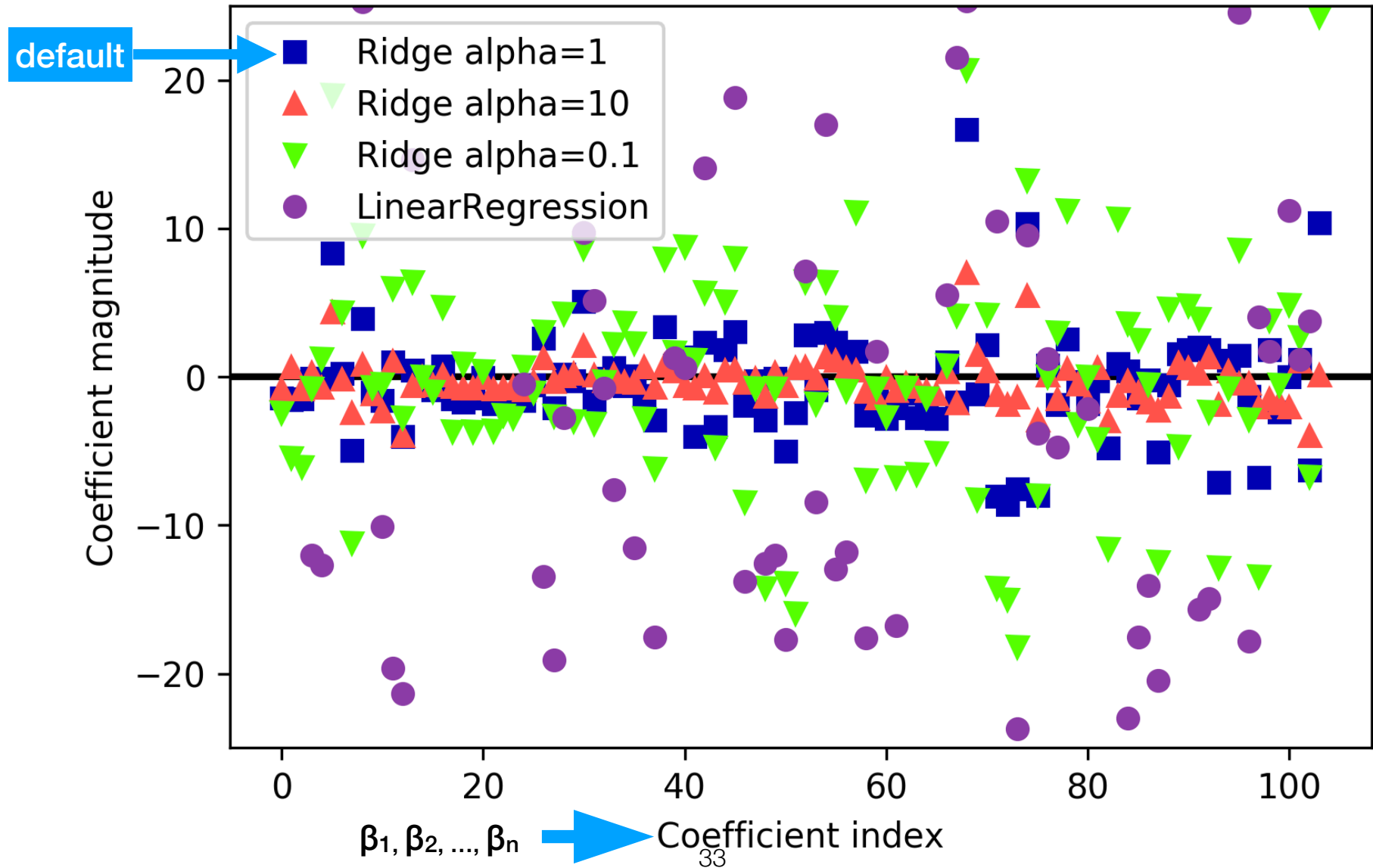
Alma mater University of Waterloo, Stanford University

Known for LASSO method

Ridge and Lasso



Ridge Regression



Lasso Regression

default



Lasso alpha=1



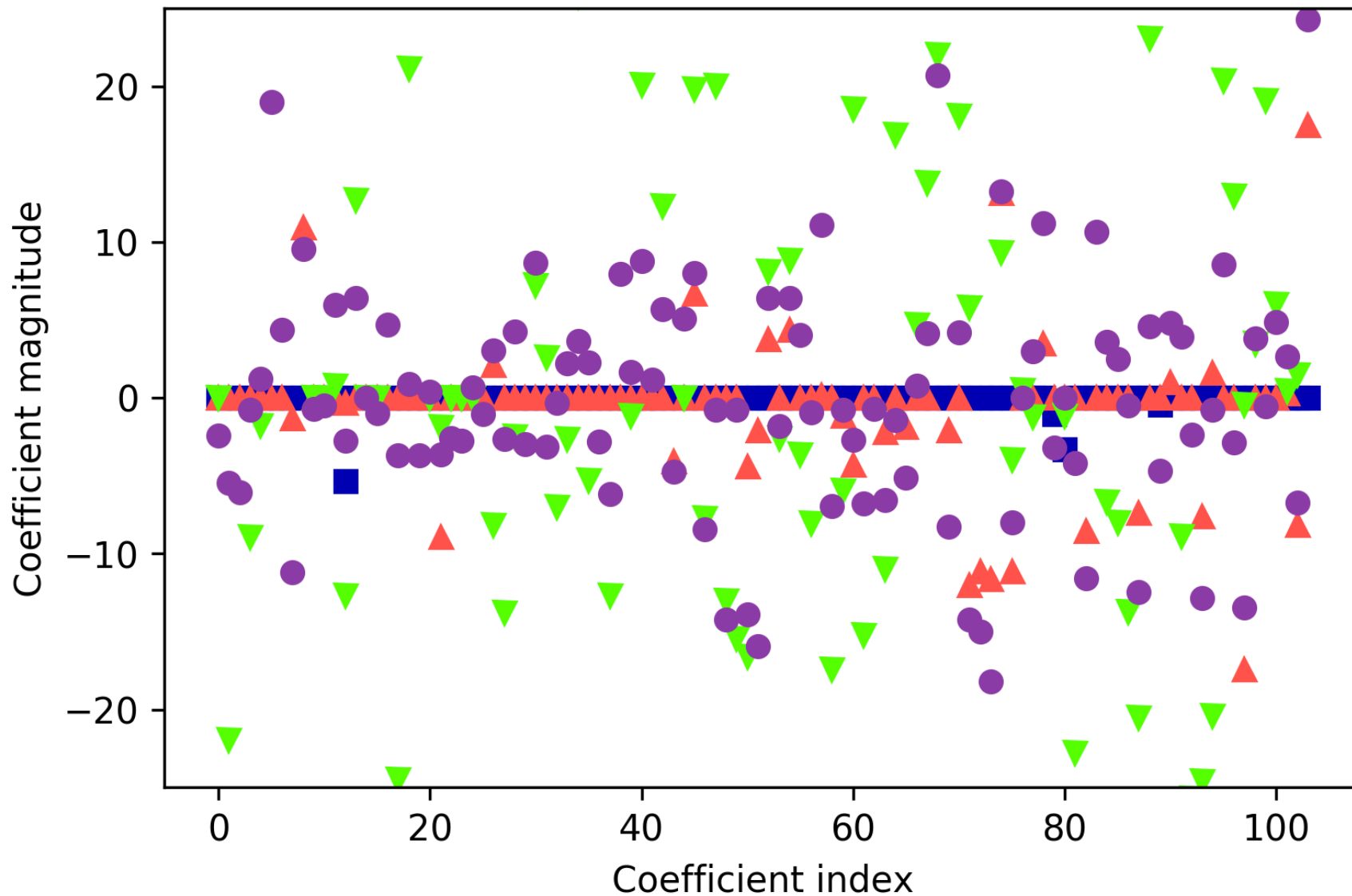
Lasso alpha=0.0001



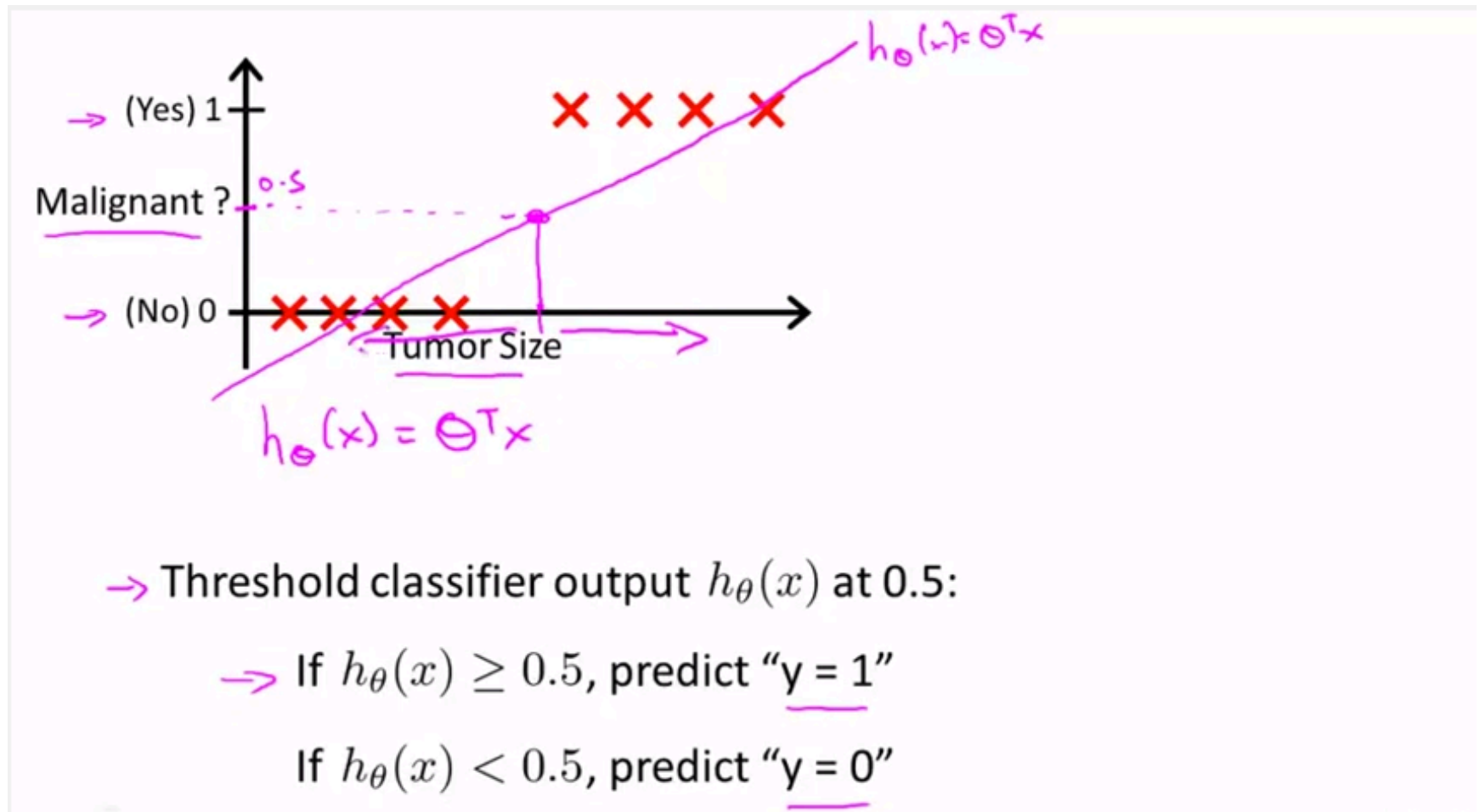
Lasso alpha=0.01



Ridge alpha=0.1



For Classification: **Don't you ever consider!** – wait until we get to log regression)



Reading

- ➡ See recommended reading provided in BB - Linear Regression Folder
- ➡ J Grus, Data Science from Scratch, ch. 14 and 15, pp.173 –188.

Tutorial

➡ See Tutorial Tasks Document