VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

AIML 231/DATA 302 – Week 6

# Regression Analysis

Dr Bach Hoai Nguyen

School of Engineering and Computer Science

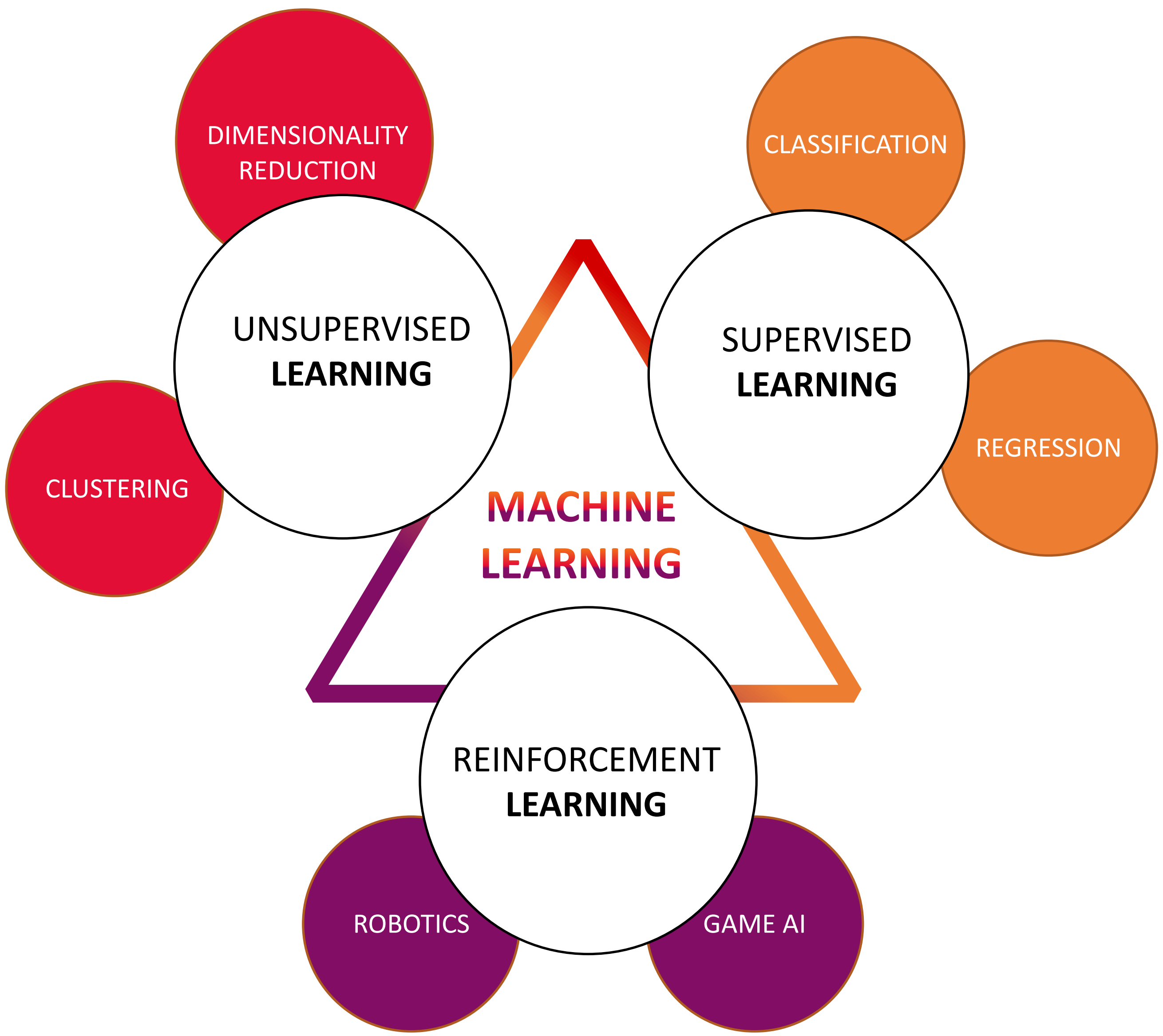
Victoria University of Wellington

Bach.Nguyen@vuw.ac.nz

# Week Overview

- Main Concepts in Regression
- Linear regression
- Regression metrics

# Regression Methods



# House price?



Dependent variables / Label / Output

| Price | Floor space | Rooms | Lot size | Appartment | Row house | Corner house | Detached |
|---|---|---|---|---|---|---|---|
| 250000 | 71 | 4 | 92 | 0 | 1 | 0 | 0 |
| 209500 | 98 | 5 | 123 | 0 | 1 | 0 | 0 |
| 349500 | 128 | 6 | 114 | 0 | 1 | 0 | 0 |
| 250000 | 86 | 4 | 98 | 0 | 1 | 0 | 0 |
| 419000 | 173 | 6 | 99 | 0 | 1 | 0 | 0 |
| 225000 | 83 | 4 | 67 | 0 | 1 | 0 | 0 |
| 549500 | 165 | 6 | 110 | 0 | 1 | 0 | 0 |
| 240000 | 71 | 4 | 78 | 0 | 1 | 0 | 0 |
| 340000 | 116 | 6 | 115 | 0 | 1 | 0 | 0 |

Independent variables / Features / Attributes/ Predictors

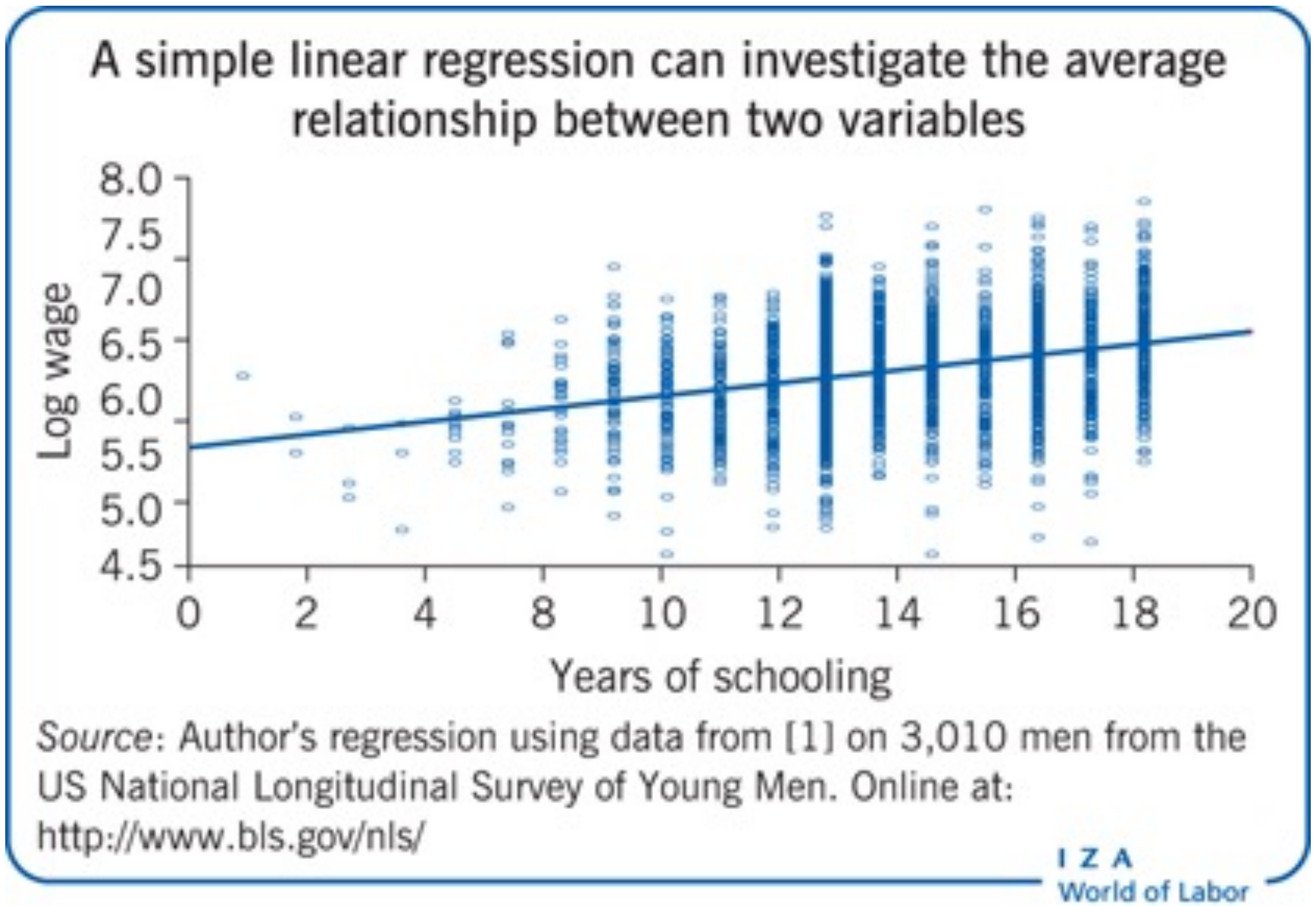
# Regression Analysis

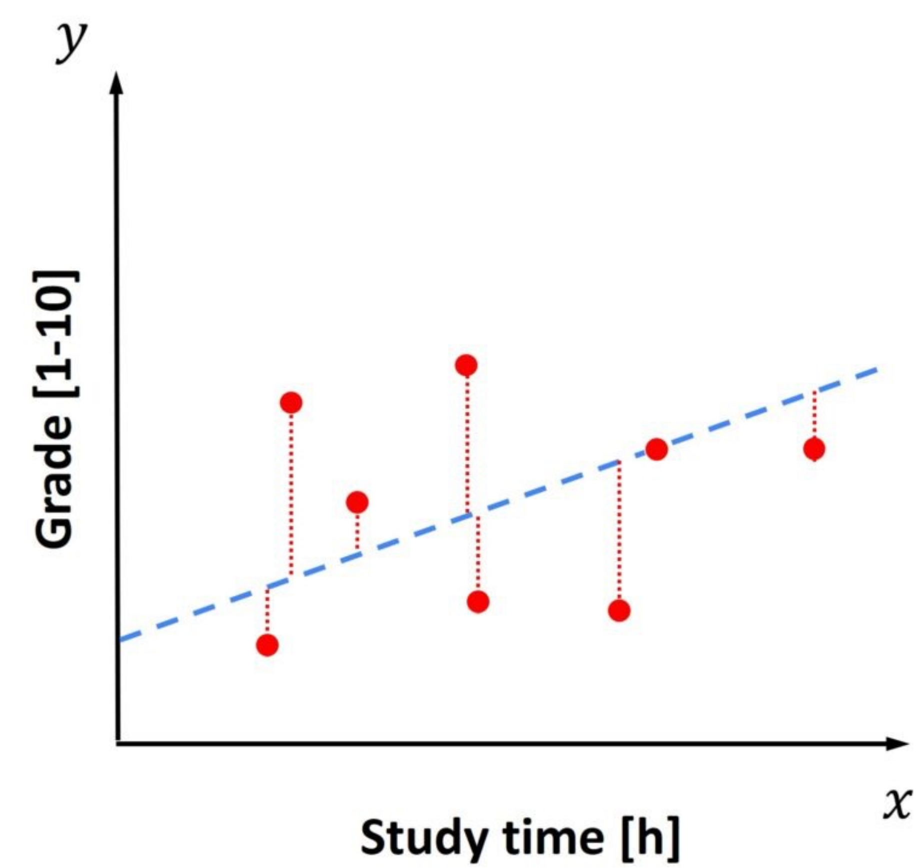
Produce a regression equation

- Regression analysis is widely used for prediction
- Describe the relationships between a set of independent variables and the dependent variable
- Describe how the changes in each independent variable *($X_i$)* are related to changes in the dependent variable (Y)
- Difference between Regression and Classification?

Output: A continuous quantify output vs. A discrete class label

# Simple Linear Regression




intercept slope error

$$y_i = w_0 + w_1\, x_i + \epsilon_i$$

Find $w_0$ and $w_1$ that minimise the total square error

$$\epsilon_1{}^2 + \epsilon_2{}^2 + \ \dots + \epsilon_n{}^2$$

$= (y_1 - w_0 - w_1\, x_1)^2 + (y_2 - w_0 - w_1\, x_2)^2 + \ \dots + (y_n - w_0 - w_1\, x_n)^2$

$= \sum_{i=1}^{N} (y_i - w_0 - w_i\, x_i)^2$

# Multiple Linear Regression

- in 1-d: fit a *straight line…*
- in more dimensions: fit a hyper*plane*
- one intercept, but many slopes, usually called ***coefficients/weights***

intercept, slopes, error

$$y_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_d x_{id} + \epsilon_i$$

$$y_i = \sum_{k=0}^{d} w_k x_{ik} + \epsilon_i$$

Find weight vector $\boldsymbol{w} = (w_0, w_1, \ldots, w_d)$ and that minimises the total square error

SquaredError = $\sum_{i=1}^{N} (y_i - \sum_{k=0}^{d} w_k x_{ik})^2$

# Regularisation

- If $\boldsymbol{w}$ is not controlled:
  - they can explode
  - hence, overfitting
- Add a penalty to control $\boldsymbol{w}$ -> regularisation

- L2 regularisation/Ridge regression

$$\text{SquaredError} + \alpha \sum_{k=1}^{d} w_k^2$$

  - shrinks the coefficients towards zero but does not set them exactly to zero

- L1 regularisation/Lasso regression

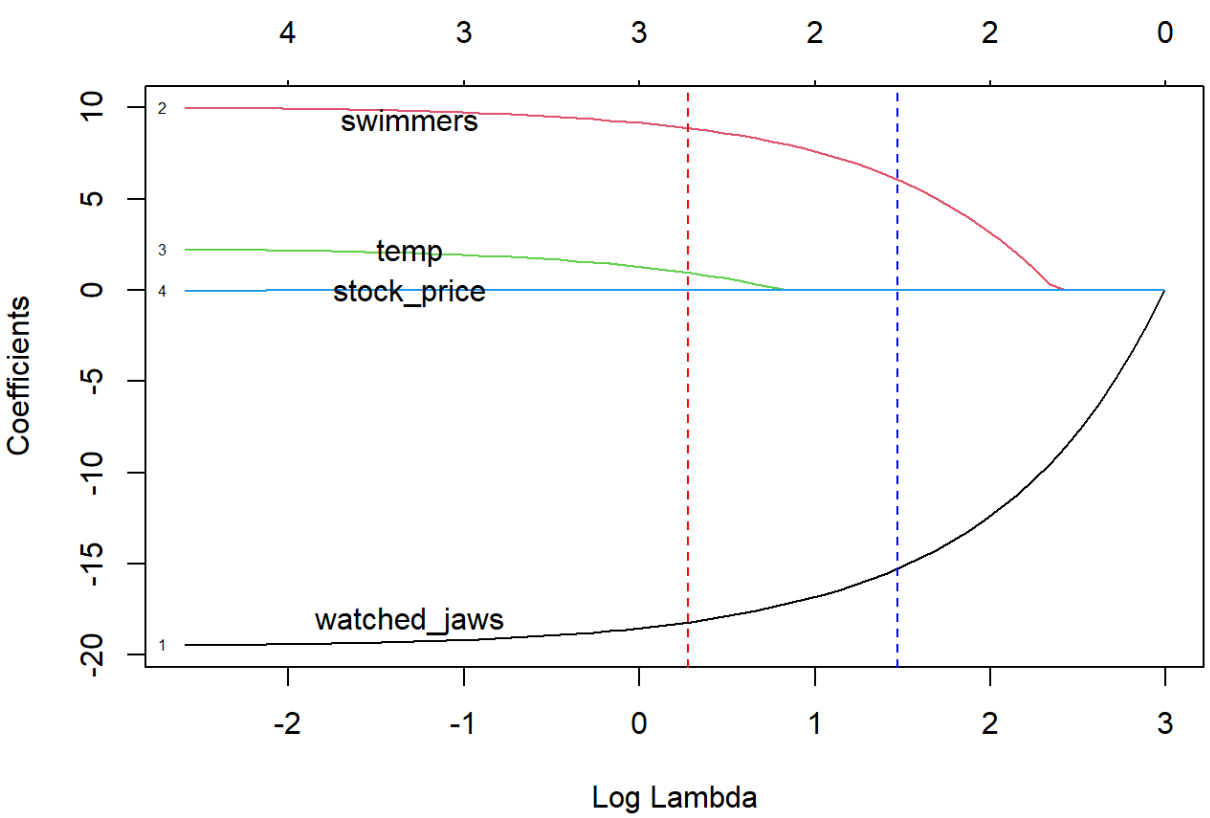
$$\text{SquaredError} + \alpha \sum_{k=1}^{d} |w_k|$$

  - setting some coefficients exactly to zero
  - effectively performing embedded feature selection

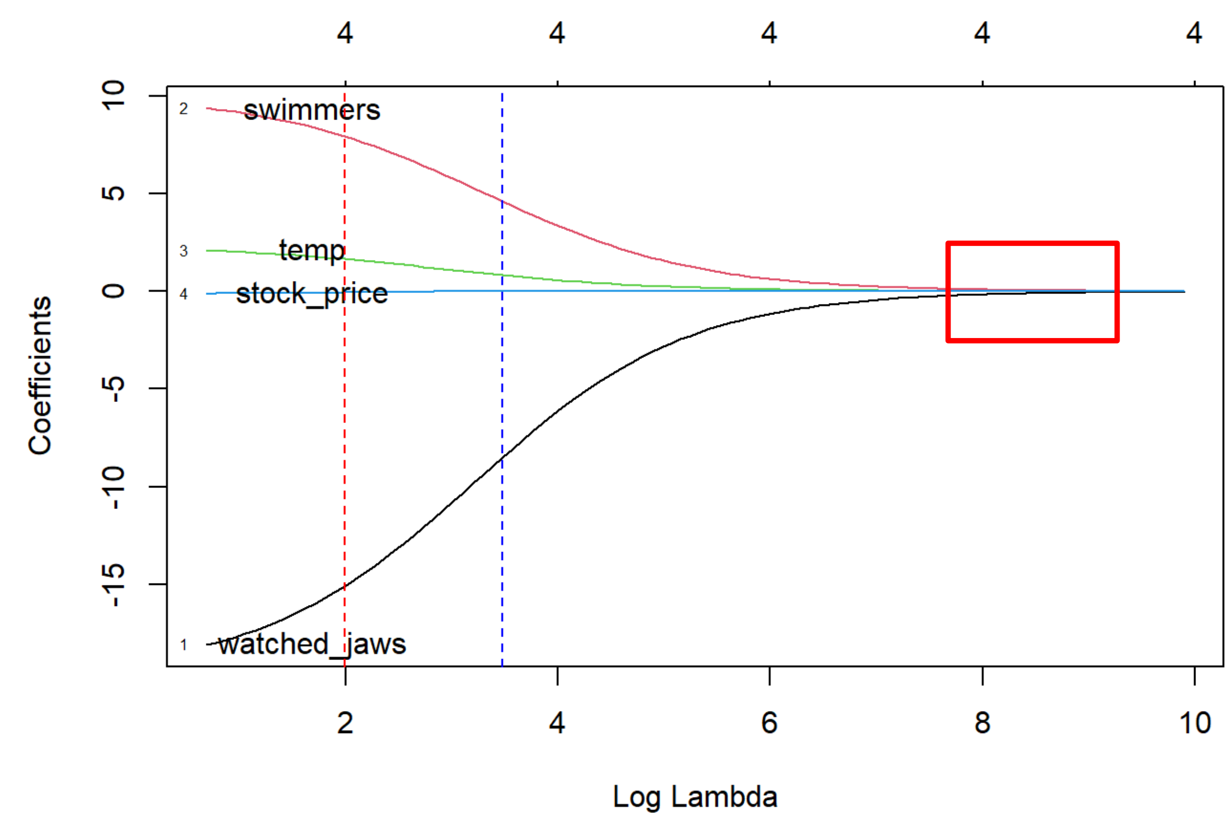
# Lasso Regression vs Ridge Regression

## Toy example: Predict number of shark attacks

- swimmers: number of swimmers
- Temp: average temperature
- watched_jaws: Percentage of swimmers who watched iconic Jaws movies
- stock_price: The price of your favourite tech stock that day (irrelevant feature)




Lasso Regression




Ridge Regression

https://www.atakanekiz.com/technical/understanding-lasso-and-ridge-regression/

# Non-linear Regression

- Polynomial Regression
- Gaussian Process Regression
- Exponential Growth Regression
- Logistic Growth Regression

….

- Genetic Programming: no model assumption 😍

# Thinking


https://becominghuman.ai/introduction-to-evaluation-metrics-e1254b81a1e8

# Mean Squared Error

- Mean Squared Error (MSE) – the most commonly used metric

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2$$

- MSE basically measure average squared error of the predictions
  - Very commonly used measure
  - If you don't have any specific preferences of the solutions to the problem
  - If you don't known any other metrics

# Root Mean Squared Error

- Root Mean Squared Error (RMSE)
  - Aims to make the scale of errors to be the same scale of target.
  - $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} = \sqrt{MSE}$

- Connection to MSE:

$$MSE(a) > MSE(b) \iff RMSE(a) > RMSE(b)$$

- Difference from MSE for gradient based methods:
  - Gradients are different: $\frac{\partial RMSE}{\partial \hat{y_i}} = \frac{1}{2\sqrt{MSE}} \frac{\partial MSE}{\partial \hat{y_i}}$
  - Travelling along MSE and RMSE is the same, but with a different learning rate, depends on MSE itself.

- Mostly, not recommended. Unless there are requirements to use it.

# Relative Squared Error

- RSE: a more interpretable measure

$$RSE = \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2}$$

  - $\overline{y}_i = \frac{1}{N}\sum_{i=1}^{N} y_i$
  - takes the total squared error and normalizes it by the total squared error of the simple predictor
  - compare between models whose errors are measured in the different units
  - should be <1 for a good model
  - R Squared / Coefficient of Determination: *1-RSE, often use for linear regression*

- Most of the time, we recommend to optimise RSE

# Mean Absolute Error

- Mean Absolute Error --- not sensitive to the outliers.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

- Compare to MSE:
  - Its penalty is smaller than that of MSE.
  - It is less sensitive to outliers in comparison to MSE.