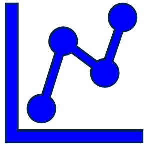
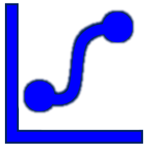
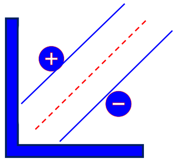
Applied Computational Intelligence

Lecture 9 – Supervised Learning

Welcome back

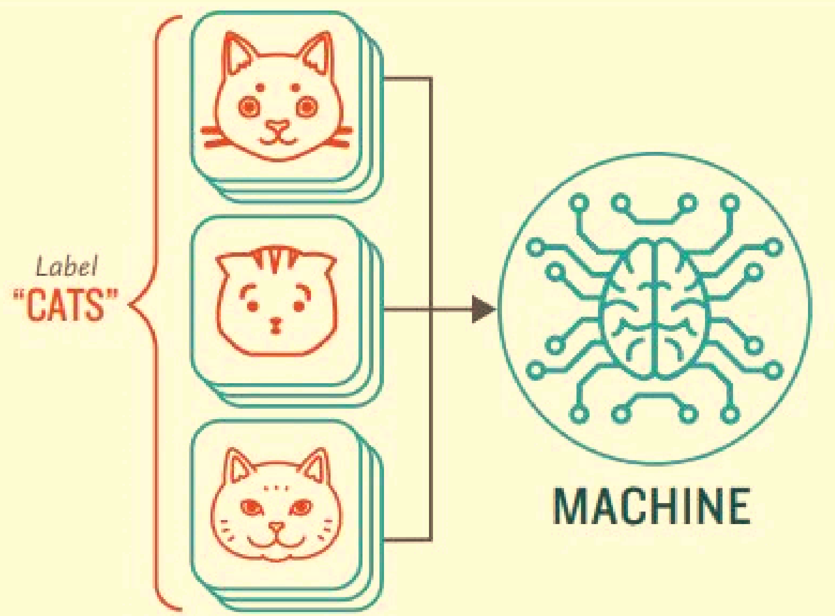
- Project Stage 2

This Session

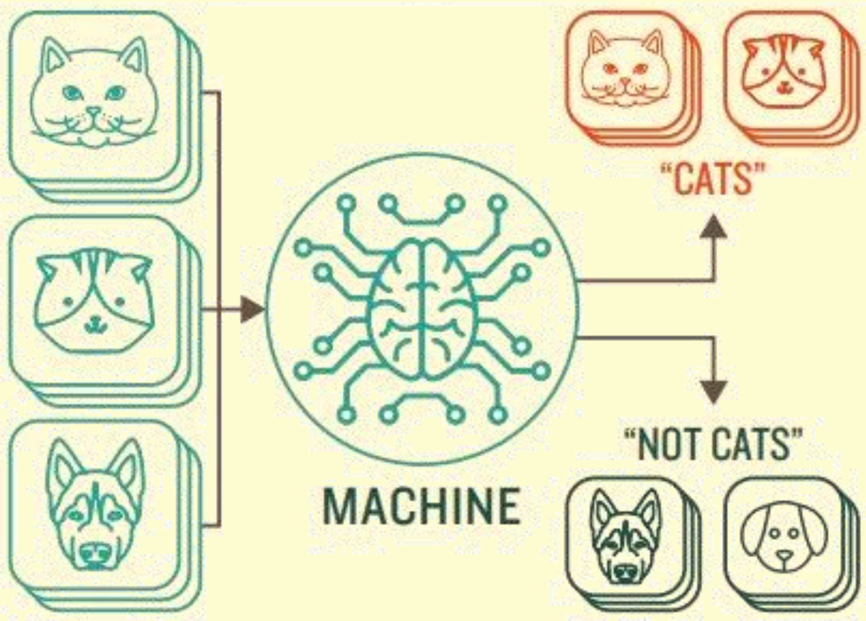
- Supervised learning
  - Linear Regression 
  - Logistic Regression 
  - Support Vector Machine 

How a Supervised ML Works

Step 1: Training with labelled I/O data

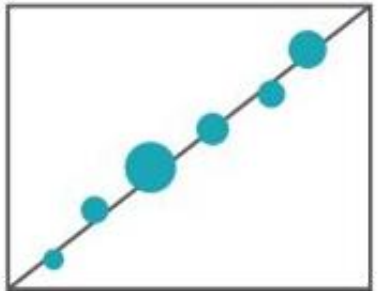


Step 2: Validate and continue refinement

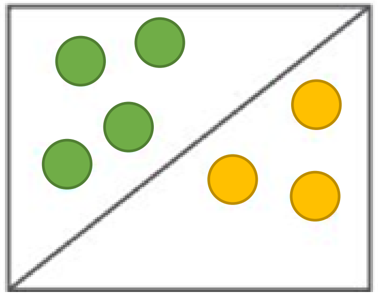


How a Supervised ML Works Cont.

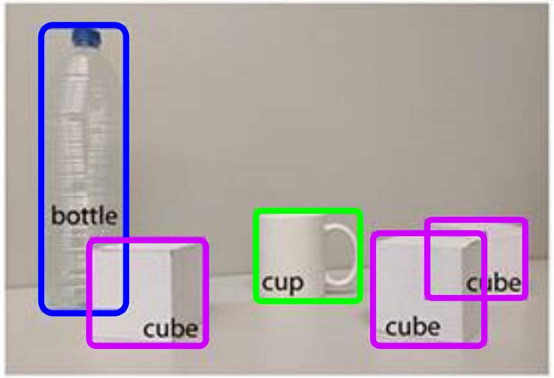
The Type of Problems to Which it's Suited



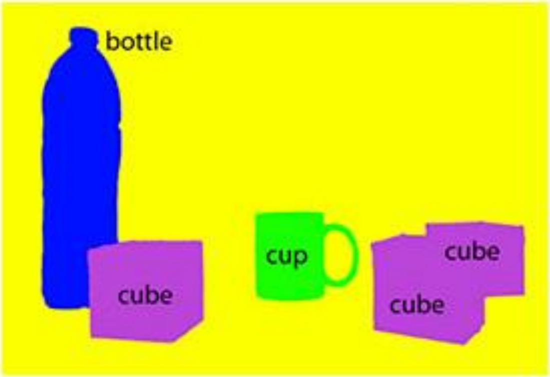
Regression
Identifying real values (weight, dollars, etc.)



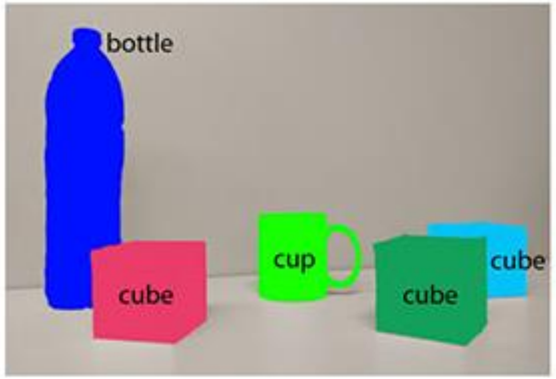
Classification
Categorizing objects into labelled groups or classes (cars, people, etc.)



Object Localization



Semantic Segmentation



Instance Segmentation

Image source: [1704.06857] A Review on Deep Learning Techniques Applied to Semantic Segmentation (arxiv.org)

Linear Regression

Overlapping content with ESOF 3675

Regression

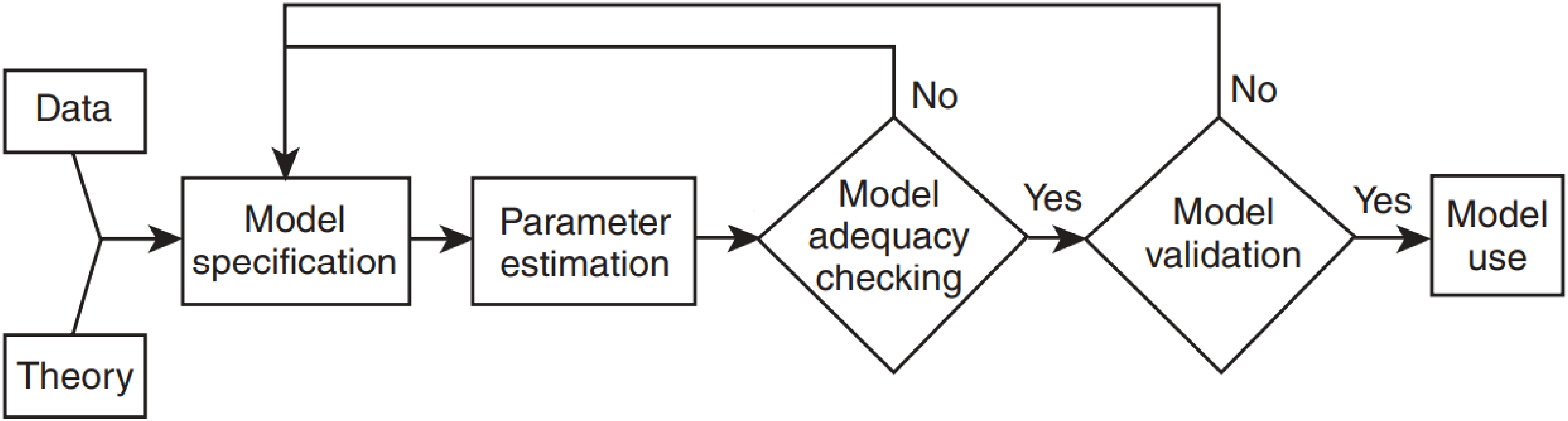
- It is a basic **statistical analysis** used **in machine learning.**
- The focus is on the **relationship between outcome**(s) and its **input variable**(s).
  - Quantify the **strength of correlation** between variables
    - ✓ how **changes** in **individual drivers affect** the outcome
  - Multiple input variables → Multivariate Regression.
- **Application:** Useful for **predicting the future values** of data based on historical information
  - E.g., Stock market prediction and trading, weather forecasting, Pilot scanning behavior analysis*
- **Note:** It is **important to remember regression** analysis **assumes that correlation relates to causation**
  - Without understanding the **context around data**, regression analysis may lead you to inaccurate predictions

* An investigation of correlation between pilot scanning behavior and workload using stepwise regression analysis, by Marvin c. Waller, Langley Research Center, National Aeronautics and Space Administration, March 1976

General Use of Regression

- Regression models are used for several purposes, including the following:
  - Data description
  - Parameter estimation
  - Prediction and estimation
  - Control

Regression - Model Building Process



Regression model - building process [3]

Linear Regression

- Used to **estimate** a **continuous** value as a **linear** (additive) **function** of other variables
  - $y := f(\textit{independent variables})$
  - E.g.,
    - `Income := f(years of education, age, gender)`
    - `House := f(median 🏠 price in neighborhood, square footage, # of bedrooms/bathrooms)`
    - `Treatment effect := f(duration of radiation, Hz of radiation, patient attributes)`

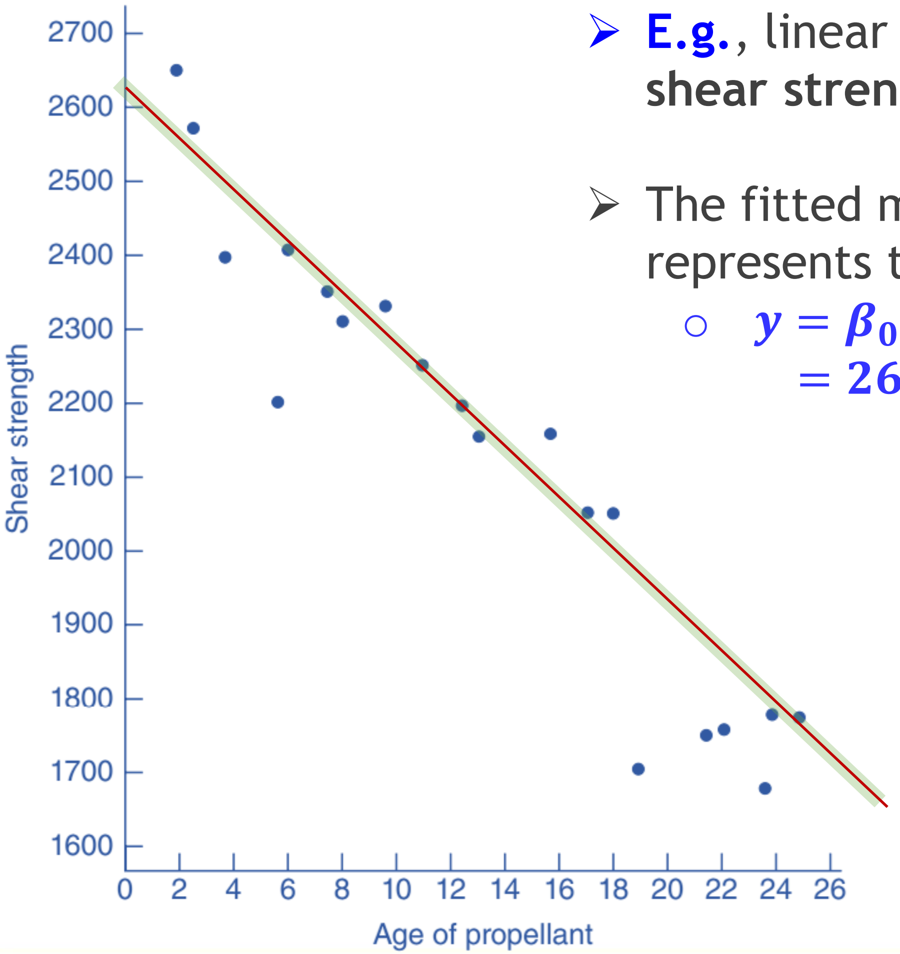
- Input: continuous or discrete

- Output:
  - A **linear expression** for predicting response as a function of drivers.
  - A **set of coefficients** that indicate the relative impact of each driver.

Linear Regression – Example

- **Model :** a single regressor, x that has a relationship with a response, y that is a straight line:
  $$y = \beta_0 + \beta_1 x$$

Obser., i	Shear Strength, y_i (psi)	Age of Propellant, x_i (weeks)
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50



- **E.g.**, linear relationship between the rocket's **shear strength and its propellant's age.**
- The fitted model (i.e., best fit line) represents the relation:
  - $y = \beta_0 + \beta_1 x$
    $= 2627.82 - 37.15x$

Linear Regression – General Model

Dataset

	F_1	F_2	…	F_p
S_1				
S_2				
$\vdots$				
S_N				

- **Assumption:** a linear relationship between the input variables and the outcome variable.
- **General model:**

$$h_{\beta}(x_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$$

$$h_{\beta}(x_i) = \beta_0 + \sum_{j=1}^{p} (x_{i,j} \cdot \beta_j) + \varepsilon$$

- The estimates for these unknown parameters are chosen so that, on average, the model provides a reasonable estimate of response variable based on the independent predictors (attributes).
- i.e., the fitted model $h_{\beta}(x)$ should minimize the overall error between the linear model and the actual observations.
- **How do we find the unknown parameters?**
  - Use an object function and an optimizer that works on minimizing or maximizing the value of the objective function

Linear Regression – General Model Cont.

- **Cost function (e.g., MSE):**

$$J(\boldsymbol{\beta}) = \frac{1}{n}\sum_{\mathrm{i=1}}^{n}\left[\boldsymbol{h}_{\boldsymbol{\beta}}(\boldsymbol{x}_{\boldsymbol{i}}) - y_i\right]^{\mathbf{2}}$$

  - The cost function can be any appropriate measures, like sum of **squared errors** (SSR) or **residual** sum of squares (RSS)).

- **Objective:** Searching for $\boldsymbol{\beta}'_j\mathrm{s}$ that produce the least value for $\boldsymbol{J}(\boldsymbol{\beta})$

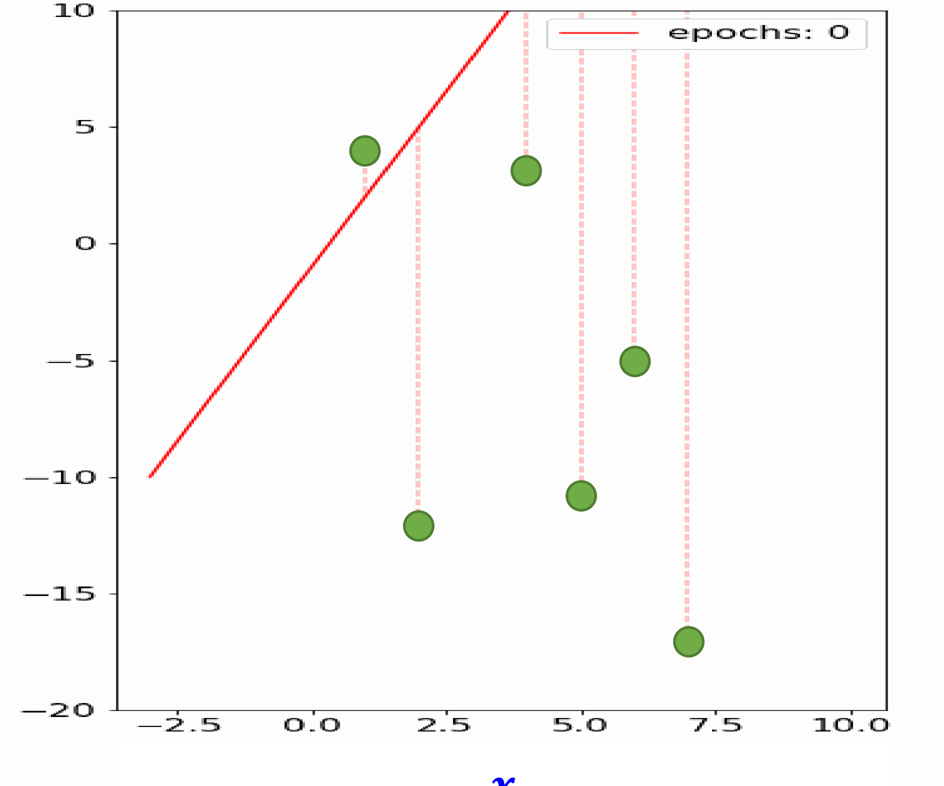
$$\boldsymbol{\beta}_j \coloneqq \min_{\boldsymbol{\beta}} \boldsymbol{J}(\boldsymbol{\beta})$$

- **How do we do this:**
  - One method is gradient decent optimizer

Linear Regression – Gradient Decent Optimizer

- **Objective:** $\boldsymbol{\beta}_j \; := \; \min_{\boldsymbol{\beta}} \boldsymbol{J}(\boldsymbol{\beta})$

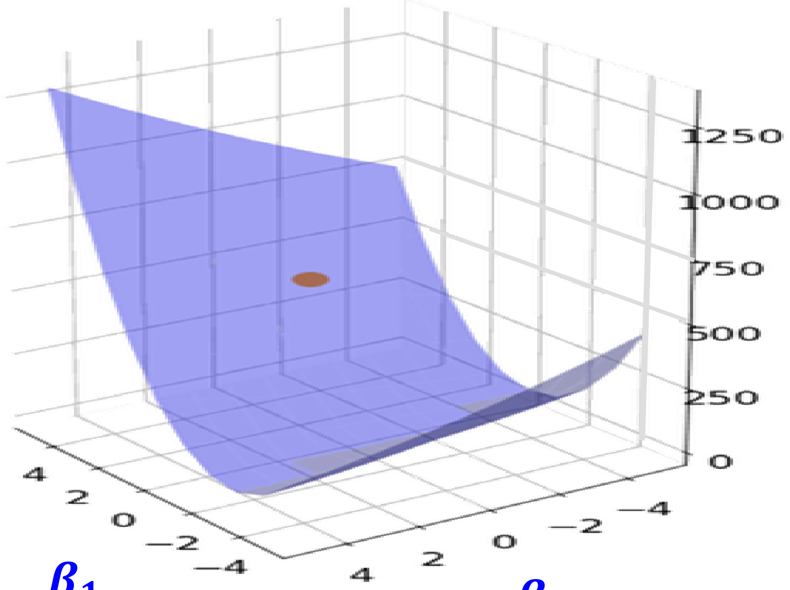
$h_{\boldsymbol{\beta}}(\boldsymbol{x})$



$\boldsymbol{x}$

Convex optimization problem

$$\boldsymbol{J}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}\left[\boldsymbol{h}_{\boldsymbol{\beta}}(\boldsymbol{x}_{\boldsymbol{i}}) - y_i\right]^2$$

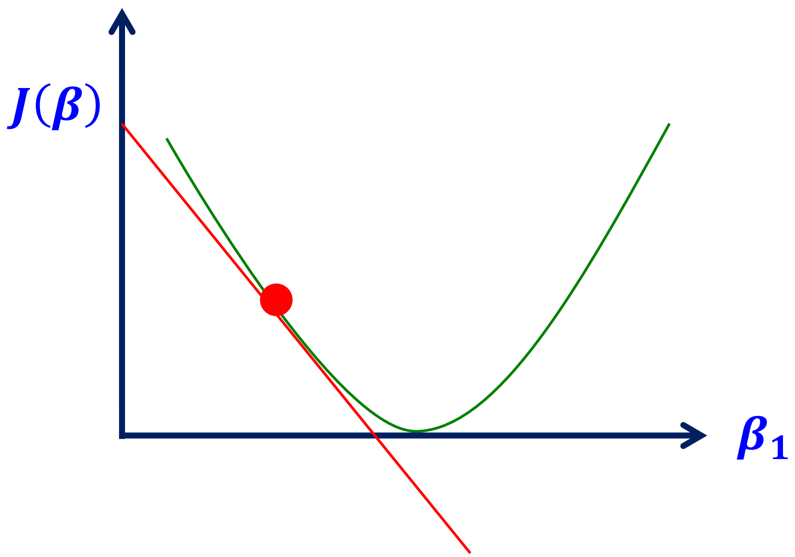


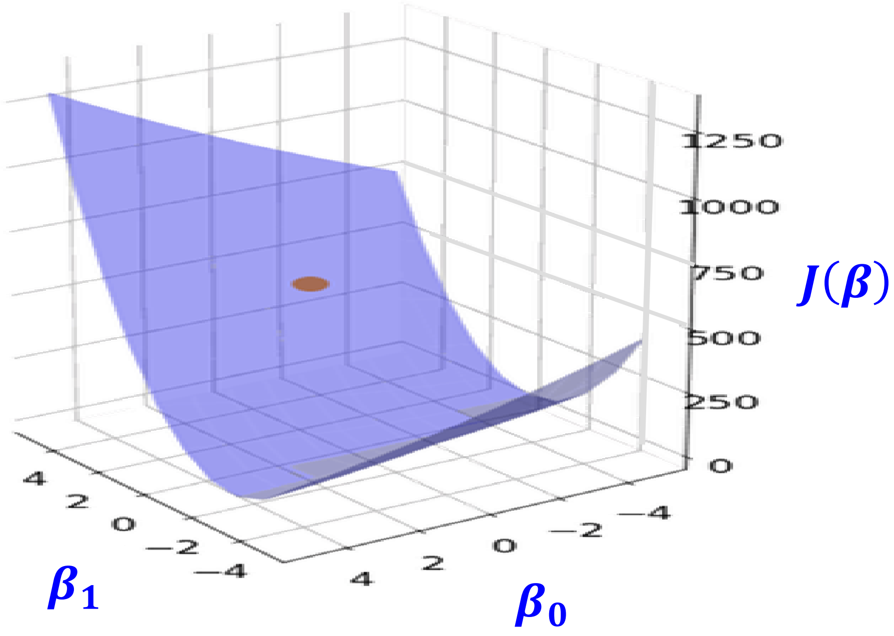
$\boldsymbol{J}(\boldsymbol{\beta})$

$\boldsymbol{\beta}_1$ $\boldsymbol{\beta}_0$

Linear Regression – Gradient Decent Optimizer Cont.

- Partial derivatives give us the slope (i.e., direction to move) in that dimension





Linear Regression – Gradient Decent Optimizer Cont.

- Partial derivatives give us the slope (i.e., direction to move) in that dimension

- Approach:
  - Pick a starting point ($\boldsymbol{\beta_0 = 0, \beta_1 = 0}$)
  - Do:
    - ✓Compute $\boldsymbol{J(\beta)}$
    - ✓Move a small amount towards decreasing loss using the derivative:
      - $\boldsymbol{\beta_0 := \beta_0 - \alpha \frac{\delta J(\beta_0, \beta_1)}{\delta \beta_0}}$
      - $\boldsymbol{\beta_1 := \beta_1 - \alpha \frac{\delta J(\beta_0, \beta_1)}{\delta \beta_1}}$

      Simultaneous update
    - ✓Repeat until convergence

Linear Regression w/ Categorical Variables

$$\text{income} = b_0 + b_1\text{age} + b_2\text{yearsOfEducation} + b_3\text{gender} + b_4\text{state}$$

- ***Gender*** is **categorical**, but **binary**
  - one variable: *Male*, which is 0 for females
- ***State*** is a **categorical** variable:
  - **50 possible values**
  - Expand it to 49 indicators (0/1) variables:
  - The remaining level is the **default level**, i.e., all indicators set to 0

```
results3 <- lm(Income~Age + Education,Gender,
    + Alabama,
    + Alaska,
    + Arizona,
    .
    .
    .
    + WestVirginia,
    + Wisconsin,
    income_input)
```

In regression, a proper way to implement a categorical variable that can take on m **different values** is to add $m - 1$ **binary variables** to the regression model.

Linear Regression - Overfitting

- Overfitting associated with **too many regression coefficients** to be estimated.

- Just adding more variables to explain a given dataset may not improve the explanatory nature of the model.

- Example:

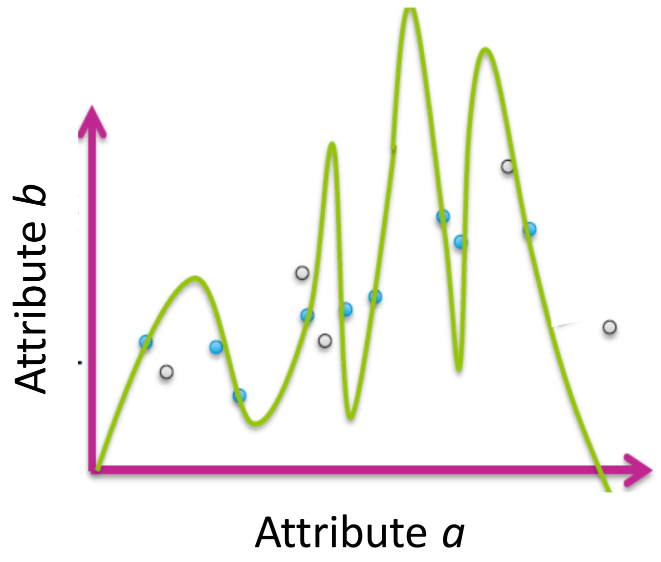
  $f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + w_3x_3$

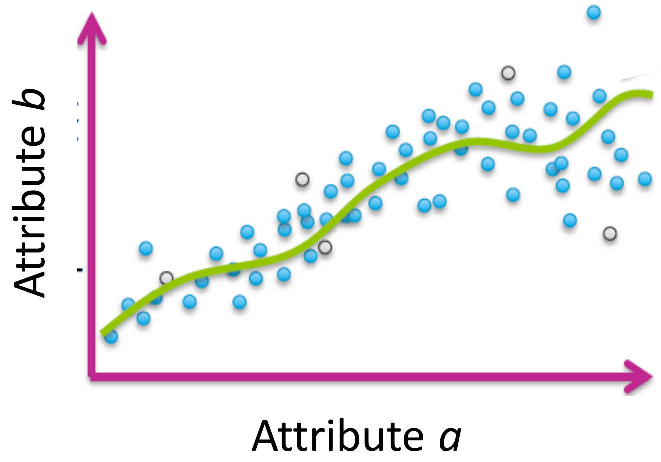
  - Let's add a fourth attribute $x_4 = x_1^2$ and add another new attribute $x_5 = \frac{x_2}{x_3}$
  - Now, the model needs to learn the parameters (weights) of the following $f(x)$.

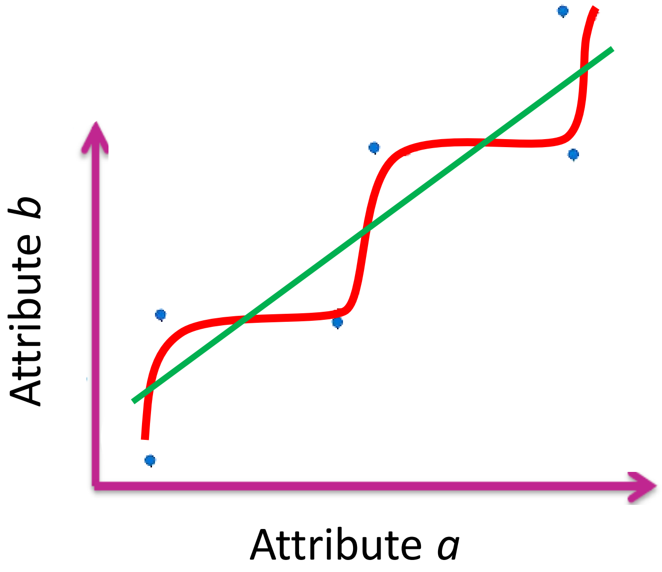
  $f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5$

  - Potentially, it can lead to overfitting and reduce model's generalizability outside the original dataset.

Linear Regression - Overfitting







- Few observations (small N)
  - ➢ rapidly overfit, as model complexity increases

- Many N (very large N)
  - ➢ harder to overfit

- Red model is overfitted, since it almost memorized all the data points.
- Green model can be an optimal solution.