Week 2

Thursday, December 30, 2021 1:37 PM

LINEAR REGRESSION

* Intro to Regression

- Types of regression models

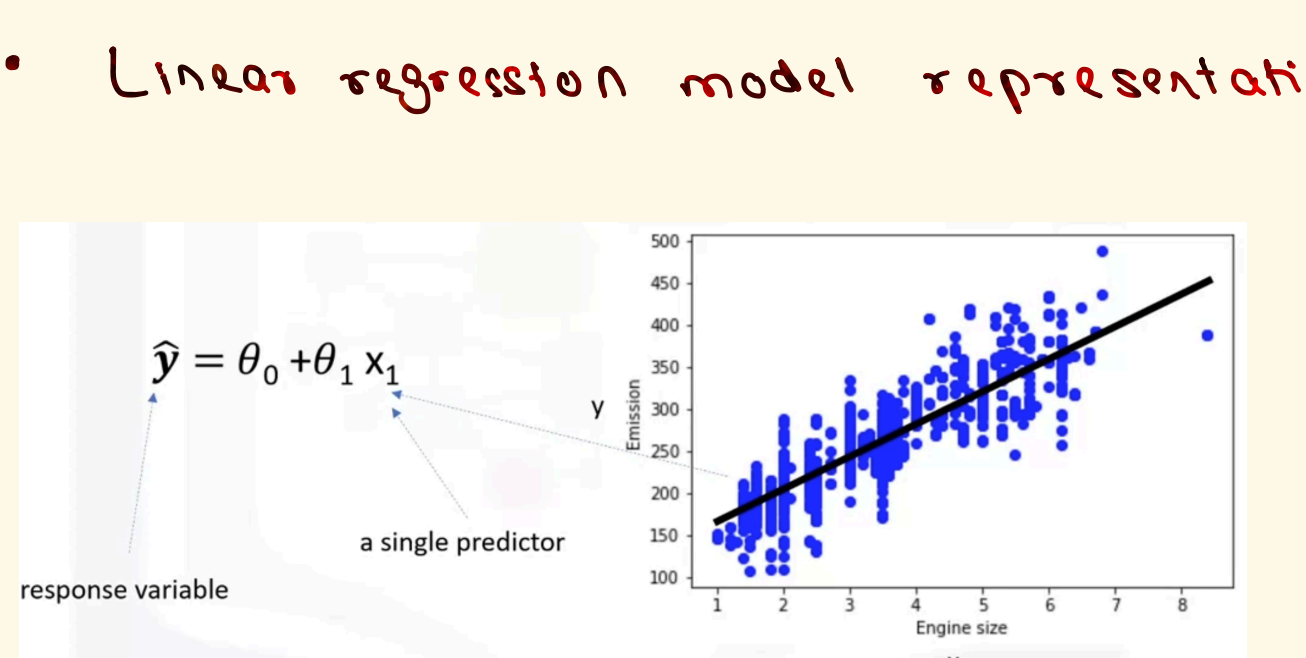
- Simple Regression:
  - Simple Linear Regression — Predict co2emission vs EngineSize of all cars
  - Simple Non-linear Regression
- Multiple Regression:
  - Multiple Linear Regression — Predict co2emission vs EngineSize and Cylinders of all cars
  - Multiple Non-linear Regression

- Applications of regression
  - sales forecasting
  - satisfaction analysis
  - Price estimation
  - Employment income

- Regression algorithms
  - Ordinal regression
  - Poisson regression (?)
  - Fast forest quantile regression (?)
  - Linear, Polynomial, Lasso, Stepwise, Ridge regression
  - Bayesian linear regression (?)
  - Neural network regression (?)
  - Decision forest regression
  - Boosted decision tree regression
  - KNN (K-nearest neighbors)

* Simple Linear Regression

- Linear regression model representation



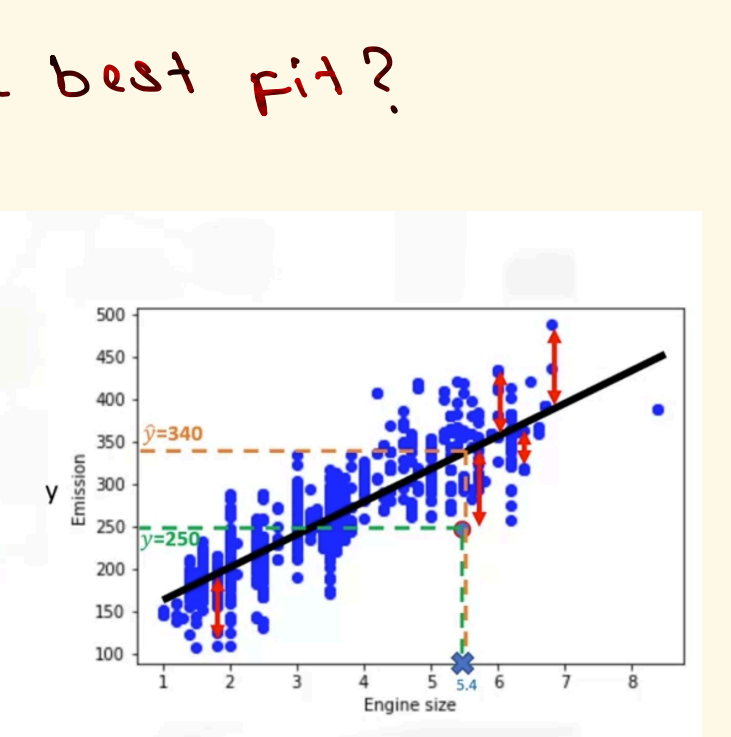
- How to find the best fit?

$x_1 = 5.4$ independent variable
$y = 250$ actual Co2 emission of x1

$\hat{y} = \theta_0 + \theta_1 x_1$
$\hat{y} = 340$ the predicted emission of x1

Error = $y - \hat{y}$ = 250 − 340 = -90

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



- Estimating the parameters

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

$\hat{y} = \theta_0 + \theta_1 x_1$

$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$

$\bar{x} = (2.0 + 2.4 + 1.5 + \ldots)/9 = 3.03$

$\bar{y} = (196 + 221 + 136 + \ldots)/9 = 226.22$

$\theta_1 = \frac{(2.0-3.03)(196-226.22)+(2.4-3.03)(221-226.22)+\ldots}{(2.0-3.03)^2+(2.4-3.03)^2+\ldots}$

$\theta_1 = 39$

$\theta_0 = \bar{y} - \theta_1\bar{x}$

$\theta_0 = 226.22 - 39 * 3.03$

$\theta_0 = 125.74$

$\hat{y} = 125.74 + 39x_1$

- Predictions with linear regression

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$\hat{y} = \theta_0 + \theta_1 x_1$

$Co2Emission = \theta_0 + \theta_1 EngineSize$

$Co2Emission = 125 + 39 EngineSize$

$Co2Emission = 125 + 39 \times 2.4$

$Co2Emission = 218.6$

- Pros of linear regression
  - Very fast
  - no parameter tuning
  - Easy to understand, and highly interpretable

* Model Evaluation in Regression Models

- Model evaluation approaches
  - Train and Test on the Same Dataset
  - Train/ Test split

- Calculating the accuracy of a model

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Test (rows 6–9), y: Actual values

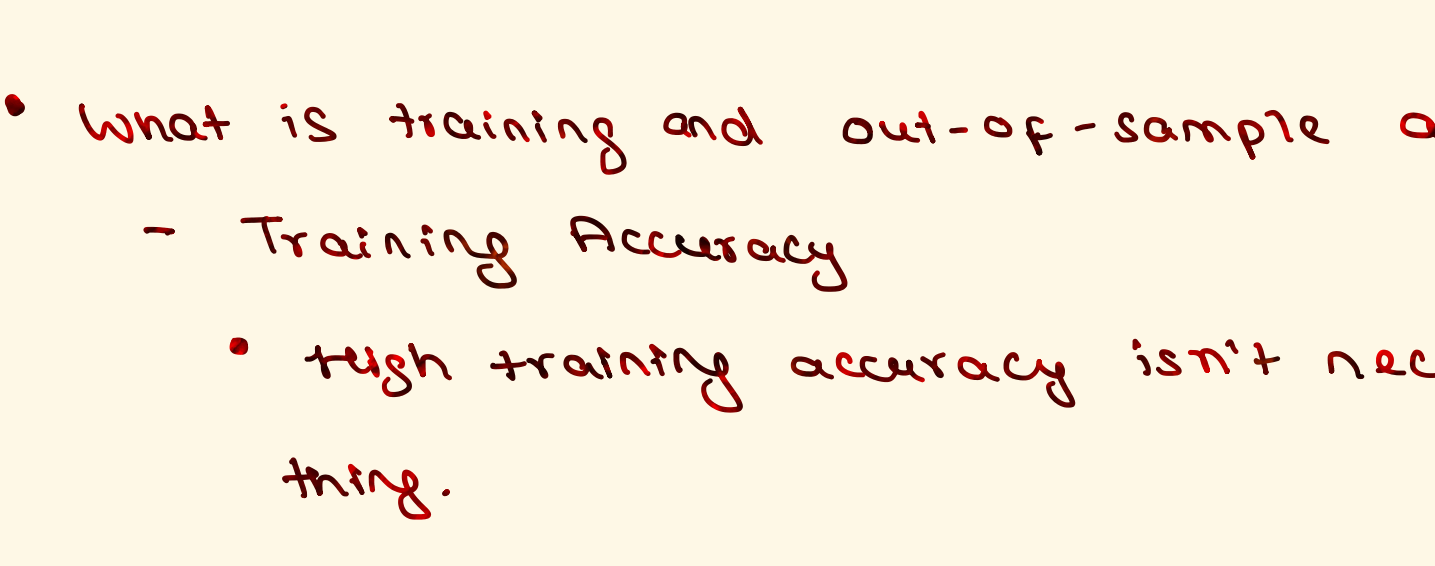
	Prediction
6	234
7	256
8	267
9	210

ŷ: Predicted values

$Error = \frac{(232-234)+(255-256)+\ldots}{4}$

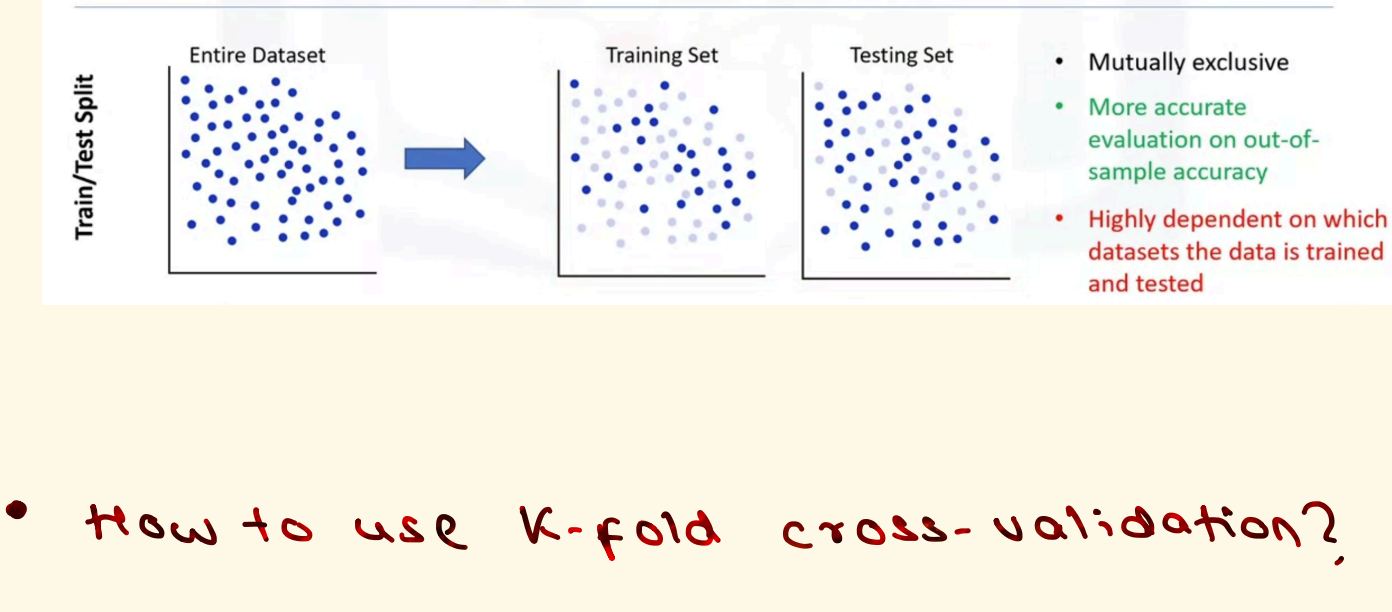
$Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$

- Train and test on the same dataset

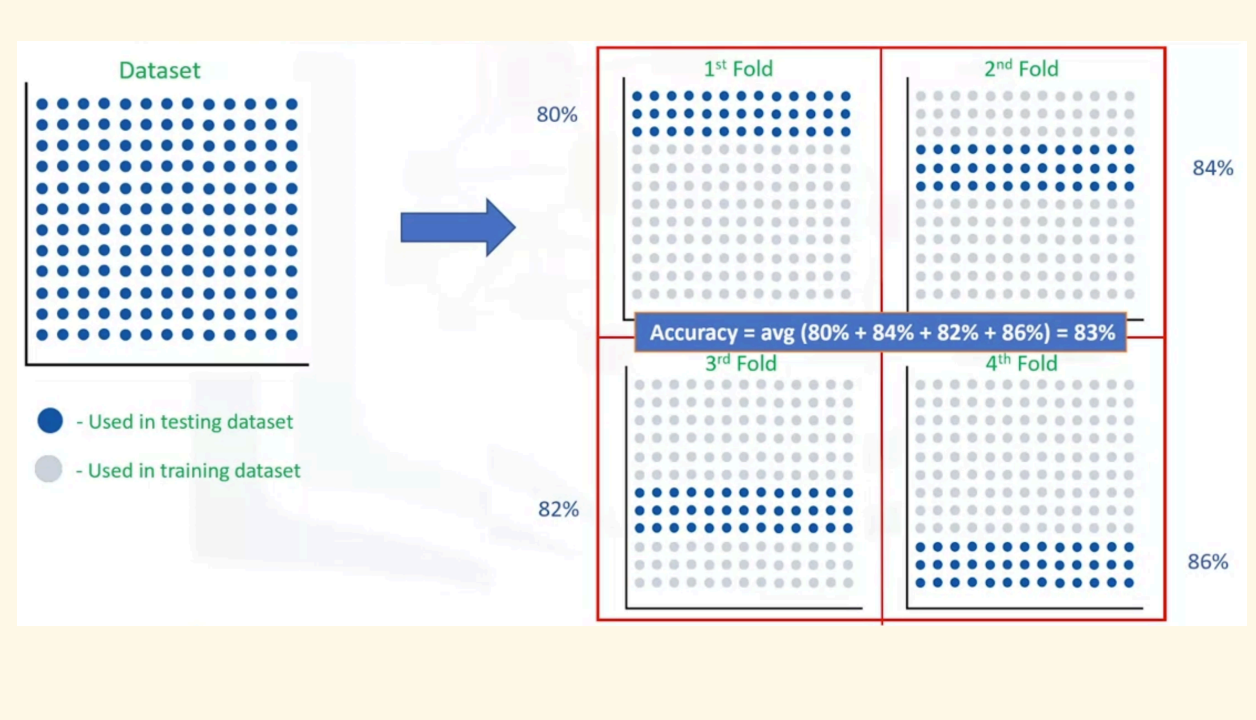


- What is training and out-of-sample accuracy?
  - Training Accuracy
    - High training accuracy isn't necessarily a good thing.
    - Result of over-fitting
      - Over-fit: the model is overly trained to the dataset, which may capture noise and produce a non-generalized model
  - Out-of-Sample Accuracy
    - It's important that our models have a high, out-of-sample accuracy
    - How can we improve oos accuracy?

- Train/Test split evaluation approach

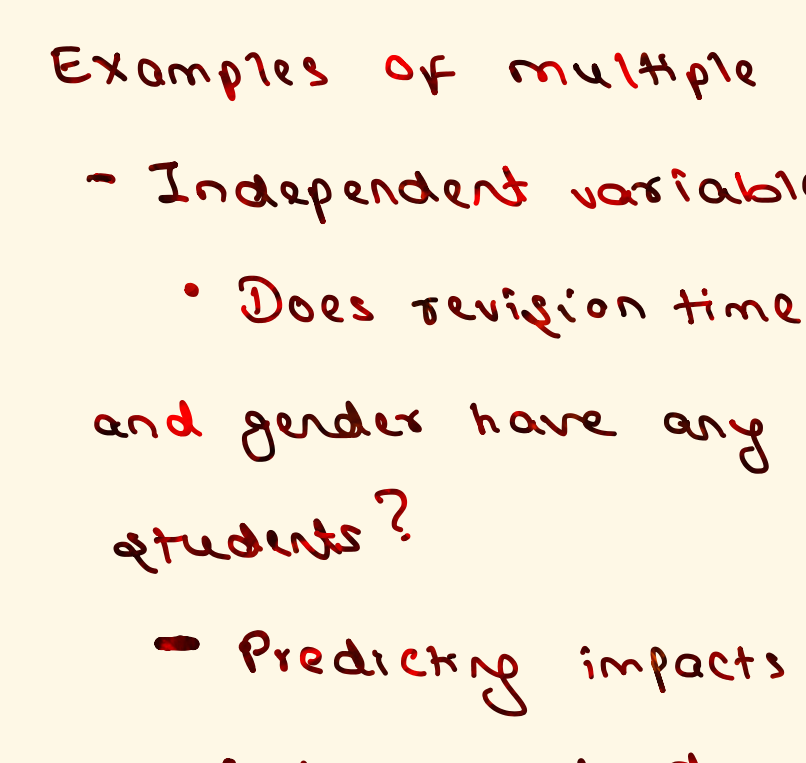


- How to use K-fold cross-validation?



* Evaluation Metrics in Regression Models

- What is an error of the model?



$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$ Mean Absolute Error

$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ Mean Square Error

$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$ Root Mean Square Error

$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$ Relative Absolute Error

$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$ Relative Square Error

$R^2 = 1 - RSE$ r^2 score

* Multiple Linear Regression

- Examples of multiple linear regression
  - Independent variables effectiveness on prediction
    - Does revision time, test anxiety, lecture attendance and gender have any effect on the exam performance of students?
  - Predicting impacts of changes
    - How much does BP go up (or down) for every unit increase (or decrease) in the BMI of a patient?

- Predicting continuous values with multiple linear regression

$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ldots$

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$

$\hat{y} = \theta^T X$ (hyperplane)

$\theta^T = [\theta_0, \theta_1, \theta_2, \ldots]$ $X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \ldots \end{bmatrix}$

X: Independent variable; Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

- Using MSE to expose the errors in the model

$\hat{y} = \theta^T X$

$\hat{y}_i = 140$ the predicted emission of x_i

$y_i = 196$ actual value of x_i

$y_i - \hat{y}_i = 196 - 140 = 56$ residual error

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

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

- Estimating multiple linear regression parameters
  - How to estimate θ?
    - Ordinary Least Squares
      - Linear algebra operations
      - Takes a long time for large datasets (10K + rows)
    - An optimization algorithm
      - Gradient Descent
      - Proper approach if you have a very large dataset

NON-LINEAR REGRESSION

* Non-Linear Regression

- What is polynomial regression?

- Some curvy data can be modeled by a polynomial regression
- For example:

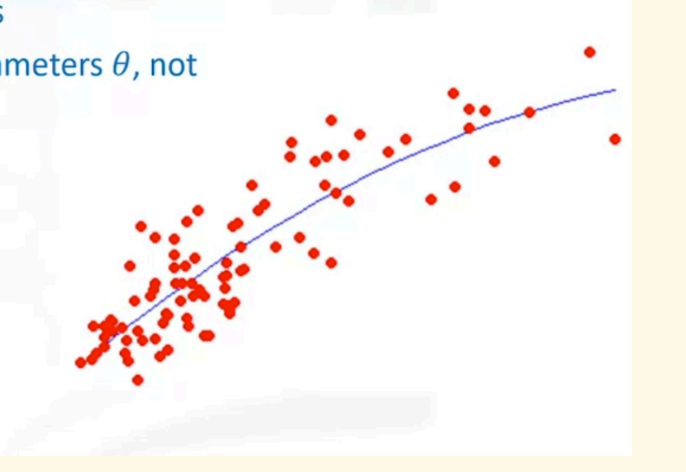
$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$

- A polynomial regression model can be transformed into linear regression model.

$x_1 = x$
$x_2 = x^2$
$x_3 = x^3$

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$ → Multiple linear regression → Least Squares

Minimizing the sum of the squares of the differences between y and ŷ



- What is non-linear regression?

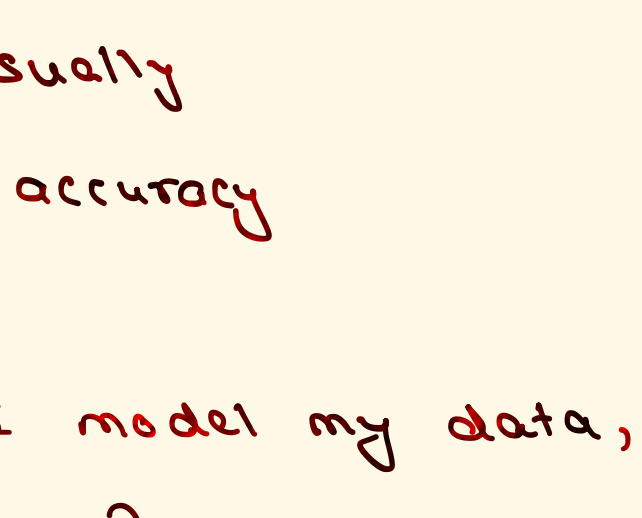
- To model non-linear relationship between the dependent variable and a set of independent variables
- ŷ must be a non-linear function of the parameters θ, not necessarily the features x

$\hat{y} = \theta_0 + \theta_2^2 x$

$\hat{y} = \theta_0 + \theta_1 \theta_2^x$

$\hat{y} = \log(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3)$

$\hat{y} = \frac{\theta_0}{1 + \theta_1^{(x - \theta_2)}}$



- Linear vs non-linear regression
  - How can I know if a problem is linear or non-linear in an easy way?
    - Inspect visually
    - Based on accuracy
  - How should I model my data, if it displays non-linear on a scatter plot?
    - Polynomial regression
    - Non-linear regression model
    - Transform your data