ARTIFICIAL INTELLIGENCE NEEDS REAL INTELLIGENCE

Regression I

Professor: Tianyuan ZHANG
tianyuan.zhang@kedgebs.com

KEDGE BUSINESS SCHOOL

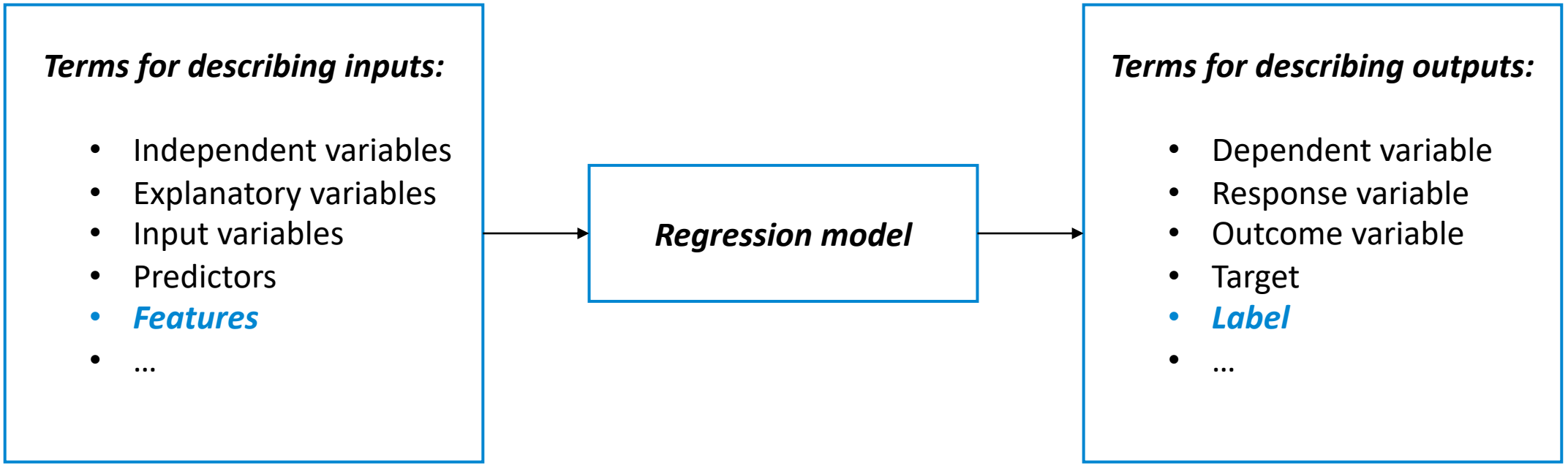
Outline

- **Introduction to Regression**
- Simple Linear Regression
- Ordinary Least Squares
- Regression Model Evaluation Metrics

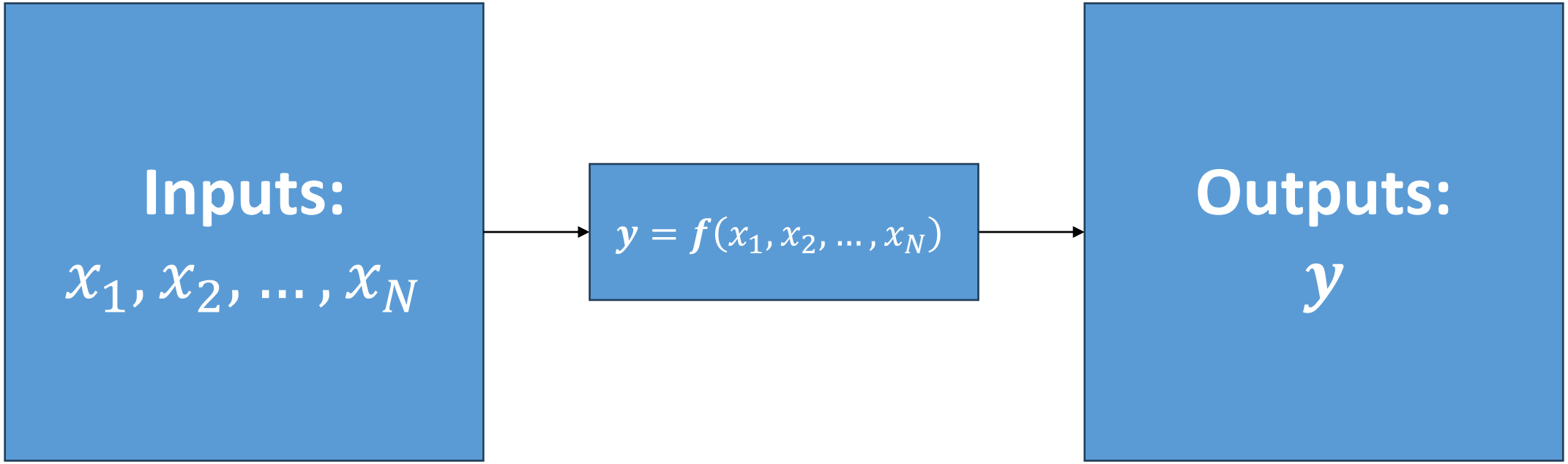
Introduction to Regression

- In statistics:
  - Regression refers to the statistical process for **estimating the relationships** between a **dependent variable** and one or more **independent variables**.
- In machine learning:
  - Regression refers to the **supervised** learning algorithms for **predicting continuous numeric values** (the label) based on the value of one or multiple predictor variables (the features).
  - Regression model needs to learn the relationships between features and label and use these relationships to make predictions on new data.

Introduction to Regression



Introduction to Regression



Introduction to Regression

- Example:
  - Given an unseen house with known size, a regression model can predict its price based on the relationship between house size and house price.

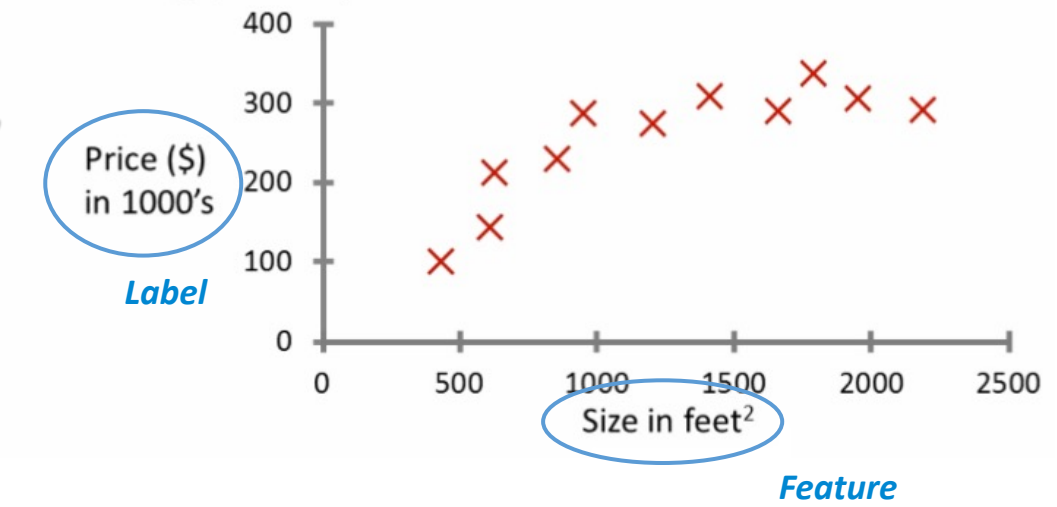


Image source: https://medium.com/@humantsattabhayya/all-about-supervised-learning-3ef9cd5efe06

Introduction to Regression

- Example:
  - Given an unseen house with known size, a regression model can predict its price based on the relationship between house size and house price.
  - The price of a house **depends** on its size.

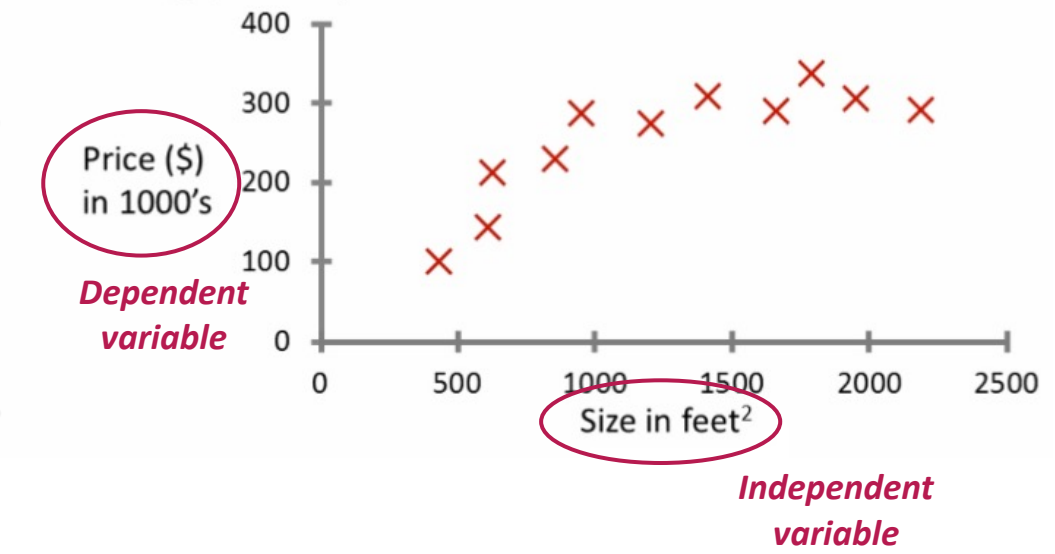


Image source: https://medium.com/@humantsattabhayya/all-about-supervised-learning-3ef9cd5efe06

Introduction to Regression

- Example:
  - Given an unseen house with known size, a regression model can predict its price based on the relationship between house size and house price.
  - The price of a house **depends** on its size.
  - The size of a house can **explain** its price.
  - If the size of a house changes, its price will change **responsively**.

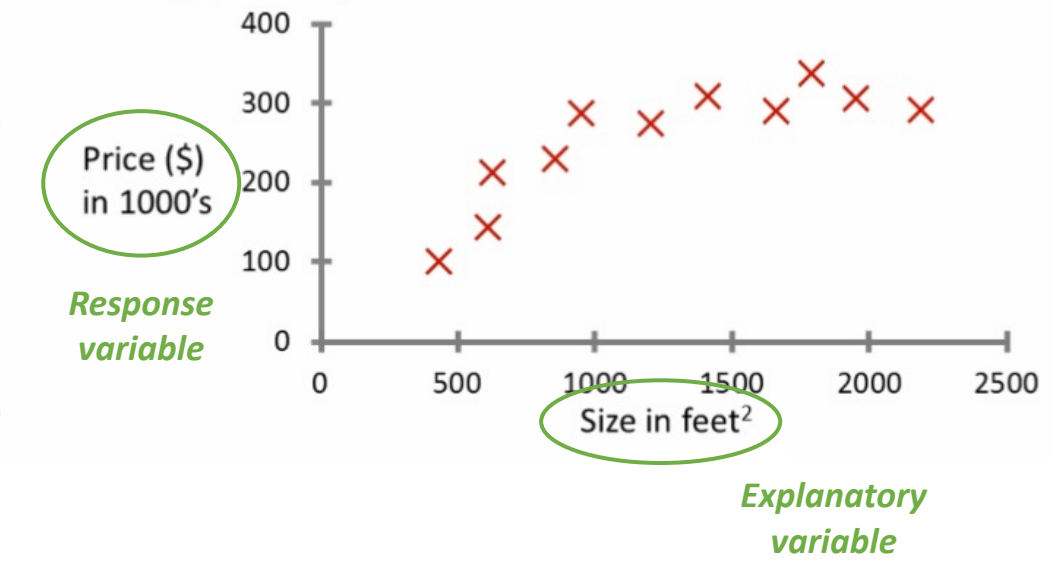


Image source: https://medium.com/@humantsattabhayya/all-about-supervised-learning-3ef9cd5efe06

Introduction to Regression

- Purpose of regression
  - To make predictions for new data points.
    - Build a regression model to estimate the lifespan of a person
      - Label: Lifespan
      - Features:
        - Gender
        - Body mass index
        - Income level
        - Average duration of sleep
        - …

Introduction to Regression

- Purpose of regression
  - To make predictions for new data points.
  - To understand and explore the relationships between variables.
    - Build a regression model to estimate the lifespan of a person
      - Label: Lifespan
      - Features:
        - Gender
        - Body mass index
        - Income level
        - Average duration of sleep
        - …

After training the regression model, you may find that the two variables are positively correlated.

Then, you may assume that increasing income is beneficial to increase lifespan.

Introduction to Regression

- Purpose of regression
  - To make predictions for new data points.
  - To understand and explore the relationships between variables.
    - Build a regression model to estimate the lifespan of a person
      - Label: Lifespan
      - Features:
        - Gender
        - Body mass index
        - Income level
        - Average duration of sleep
        - …

After training, you may find the resulting model ignores a specific feature.

Then, you may assume that this specific feature has no impact on the lifespan, or the impact is not significant.

You can try to eliminate this feature, retrain the model to see if you can have a better model.

Introduction to Regression

- Example functions powered by regression in apps:
  - Price prediction in ride-sharing apps
    - Apps like Uber and Lyft use regression models to estimate the cost of a ride. These models consider various features like distance, expected traffic, ride demand, and local fare rates to provide users with a price estimate before they book a ride.

Introduction to Regression

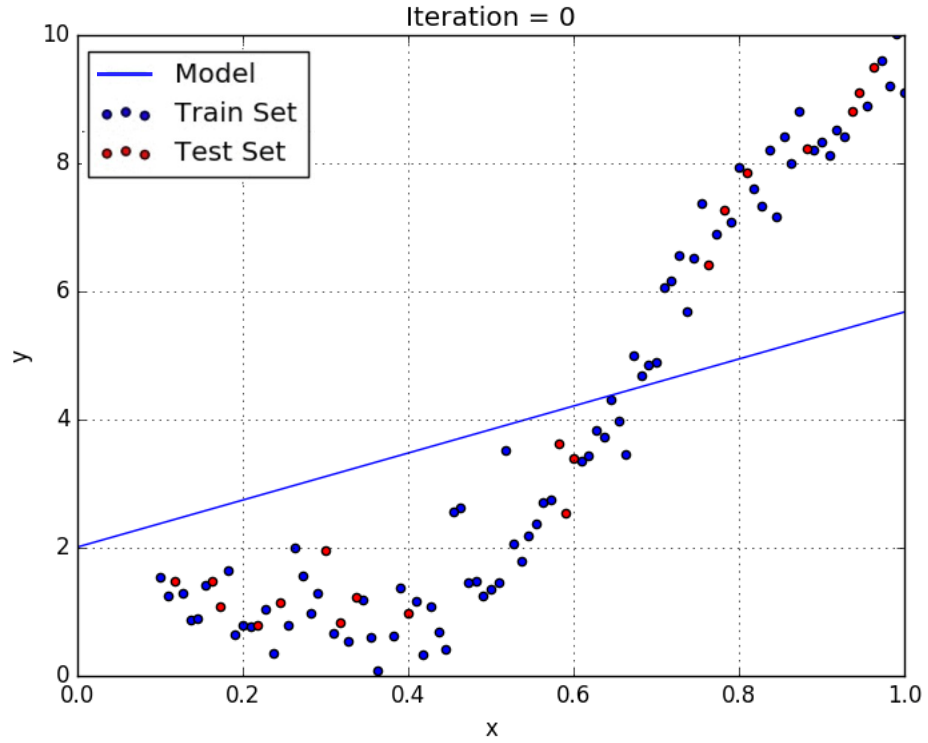
- Example functions powered by regression in apps:
  - Health tracking in fitness apps
    - Fitness apps such as MyFitnessPal or Fitbit use regression models to estimate calories burned during exercise. Inputs may include the user's weight, age, heart rate, and the type and duration of physical activity.

Introduction to Regression

- Types of regression
  - Linear regression
    - Estimate the linear relationship between feature variables and the label variable
  - Polynomial regression
    - Extend linear regression with polynomial terms to model non-linear relationships
  - Ridge regression
    - Address multicollinearity in linear regression models
  - Lasso regression
    - Perform feature selection to improve model simplicity

Introduction to Regression

- Linear regression vs. Polynomial regression



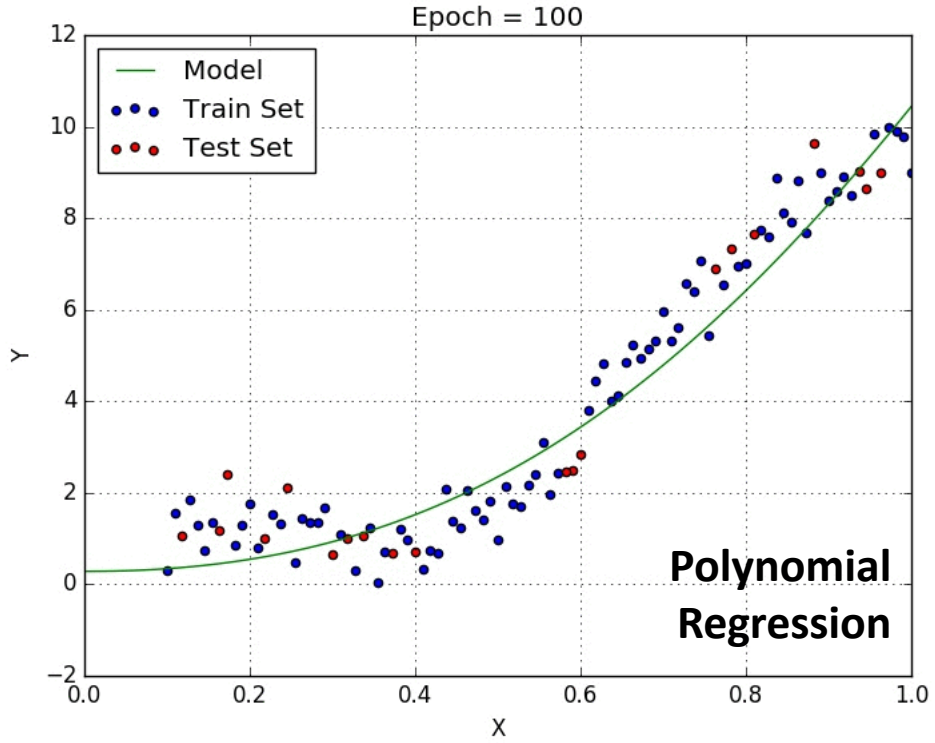


Image source: https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb

Outline

- Introduction to Regression
- **Simple Linear Regression**
- Ordinary Least Squares
- Regression Model Evaluation Metrics

Simple Linear Regression

- Linear Regression (LR)
  - Estimate the relationship between a dependent variable (y) and the independent variables $(x_1, x_2, \dots, x_N)$ by fitting a linear equation to the training dataset.
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N$
- Simple Linear Regression (SLR)
  - Linear regression with a single independent variable $(x_1 \text{ or } x)$
  - $y = \beta_0 + \beta_1 x$

Simple Linear Regression

- Simple Linear Regression (SLR)
  - $y = \beta_0 + \beta_1 x$
    - y is the dependent variable, the label we're trying to predict.
    - x is the independent variable, the single input feature.
    - β_1 is the slope of the line.
    - β_0 is the y – intercept.

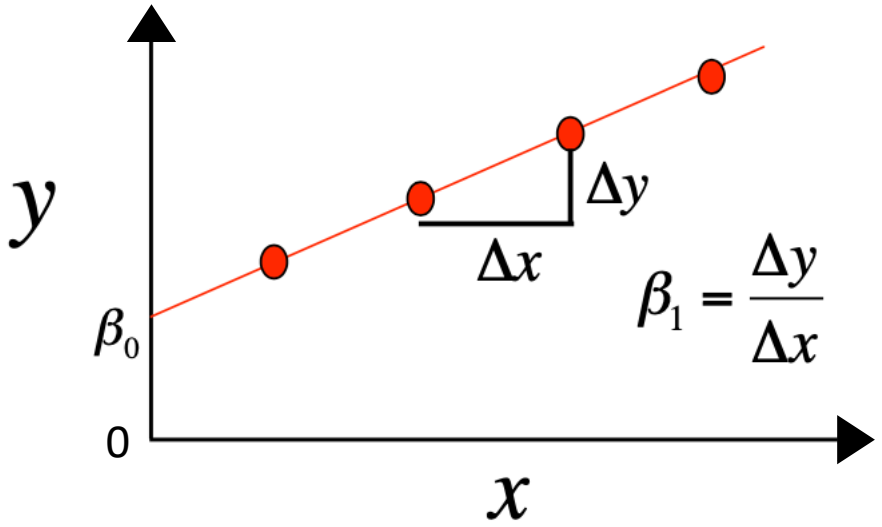
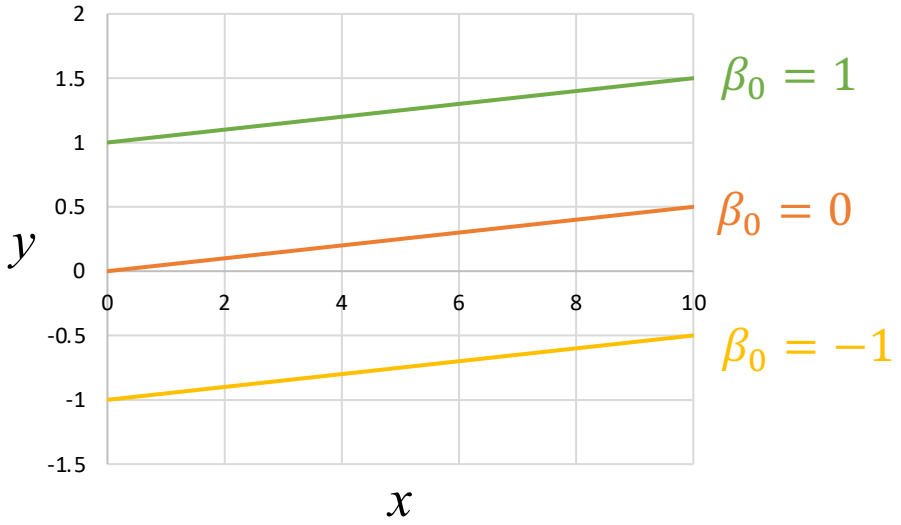


Image source: https://homes.cs.washington.edu/~suinlee/genome560/lecture8.pdf

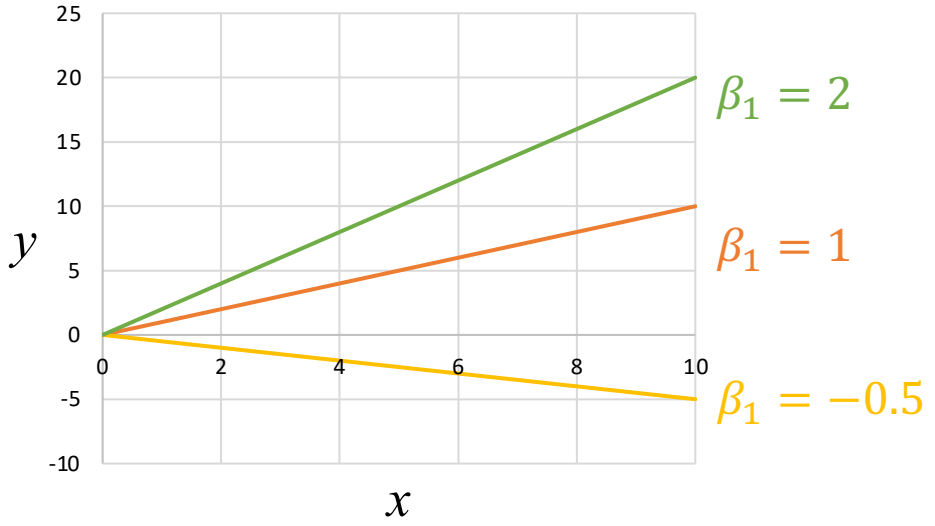
Simple Linear Regression

- Simple Linear Regression (SLR)
  - β_0 is the y – intercept.
    - The y – intercept describe the point where the regression line crosses the y – axis.
    - The value of the dependent variable (DV) when the independent variable (IV) is equal to zero.



Simple Linear Regression

- Simple Linear Regression (SLR)
  - β_1 is the slope of the line.
    - The slope describes both the **direction** and **steepness** of the line.
    - In SLR. the slope describes how the dependent variable (DV) changes in response to changes in the independent variable (IV).
      - Positive → Direct relationship → As the IV increases, the DV also increases.
      - Negative → Inverse relationship → As the IV increases, the DV decreases.
      - Steep → Strong relationship → A small change in the IV results in a large change in the DV.
      - Shallow → Weak relationship → The IV must change significantly to have a noticeable effect on the DV.



Simple Linear Regression

- Simple Linear Regression (SLR)
  - $y = \beta_0 + \beta_1 x$
  - The goal of training a SLR model is to find the line (defined by β_0 and β_1) that **'best fits'** the training dataset.
  - What does **'best fit'** mean?

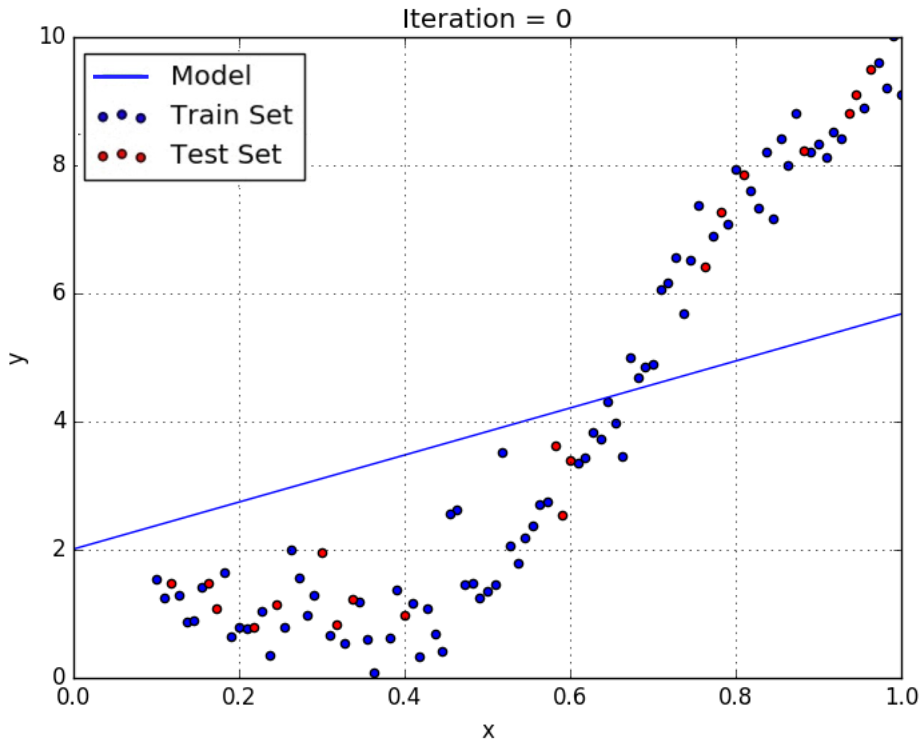
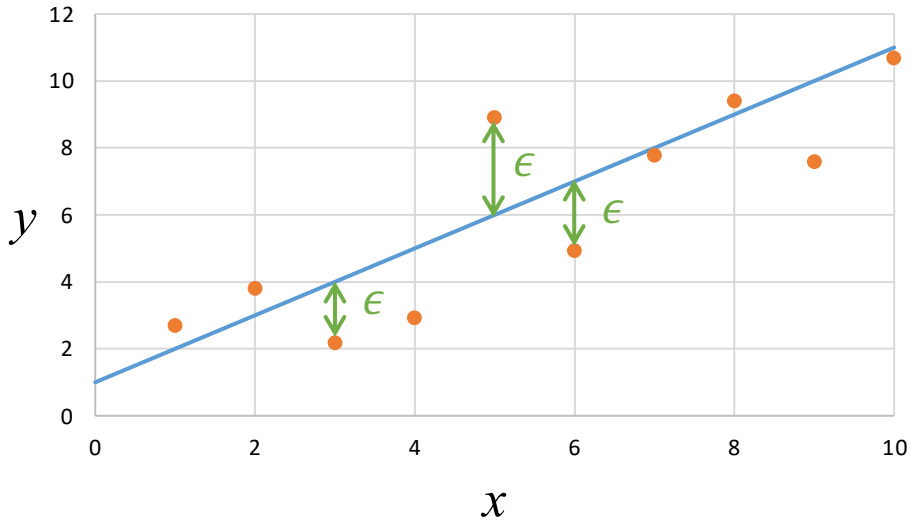


Image source: https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb

Simple Linear Regression

- Simple Linear Regression (SLR)
  - $y = \beta_0 + \beta_1 x$
  - For each data point (x_i, y_i), the SLR model can make a prediction $\hat{y}_i$
    - $\hat{y}_i = \beta_0 + \beta_1 x_i$
  - The prediction $\hat{y}_i$ may not be equal to the label value y_i, the difference between the two is called **the residual ϵ**
    - $\epsilon_i = y_i - \hat{y}_i$



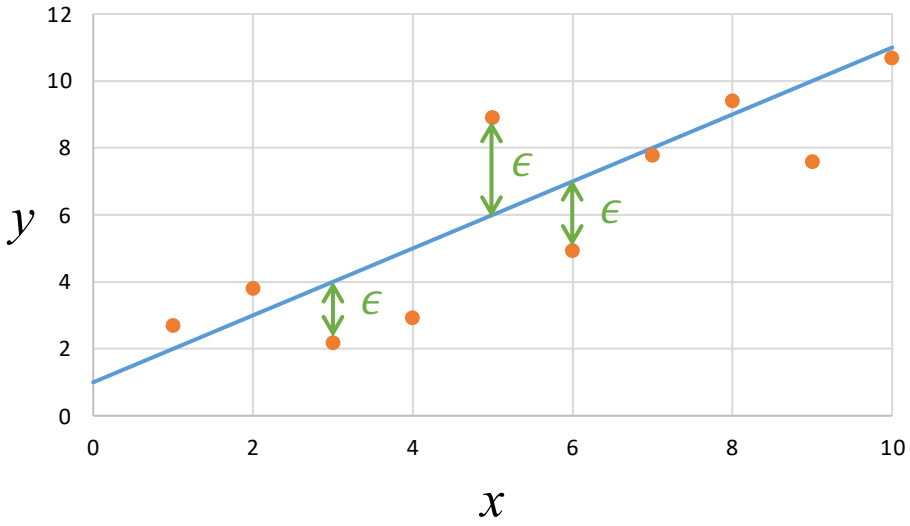
- The goal of training a SLR model is to find the best β_0 and β_1 that **minimize the residuals for all data points → 'best fits' the data points.**

Outline

- Introduction to Regression
- Simple Linear Regression
- **Ordinary Least Squares**
- Regression Model Evaluation Metrics

Ordinary Least Squares

- Simple Linear Regression $y = \beta_0 + \beta_1 x$
  - For each data point (x_i, y_i), the SLR model can make a prediction $\hat{y}_i$
    - $\hat{y}_i = \beta_0 + \beta_1 x_i$
  - The residual ϵ between the prediction $\hat{y}_i$ and the label value y_i
    - $\epsilon_i = y_i - \hat{y}_i$
  - The goal is to minimize the residuals for all data points, but **how?**



    - As the line moves, some points get closer to the line, and others get father from the line.
    - As the line moves, the residuals become smaller at some points and larger at others.

Ordinary Least Squares

- The ordinary least squares approach is a method to estimate the unknown parameters (β_0 and β_1) in a linear regression model by the principle of **least squares**:
  - Minimizing the sum of the squares of the residuals
    - Find $\min_{\beta_0,\beta_1} Q(\beta_0, \beta_1)$, for $Q(\beta_0, \beta_1) = \sum_i \epsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$
    - $\sum_i \epsilon_i^2$ is also called the **Sum of Squared Errors (SSE)**
  - The estimated parameters by the ordinary least squares approach is denoted as $\hat{\beta}_0$ and $\hat{\beta}_1$

Ordinary Least Squares

- Using calculus or linear algebra (details in the complementary materials on Learn), we can find:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$\bar{x} = \frac{1}{n}\sum x_i, \quad \bar{y} = \frac{1}{n}\sum y_i, \quad s_{xx} = \frac{1}{(n-1)}\sum (x_i - \bar{x})^2, \quad s_{xy} = \frac{1}{(n-1)}\sum (x_i - \bar{x})(y_i - \bar{y}).$$

Here $\bar{x}$ is the sample mean of x, $\bar{y}$ is the sample mean of y, s_{xx} is the sample variance of x, and s_{xy} is the sample covariance of x and y.

Image source: https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2014/4165bb0ea4e716909c1fa78eec116ae7_MIT18_05S14_Reading25.pdf

Outline

- Introduction to Regression
- Simple Linear Regression
- Ordinary Least Squares
- **Regression Model Evaluation Metrics**

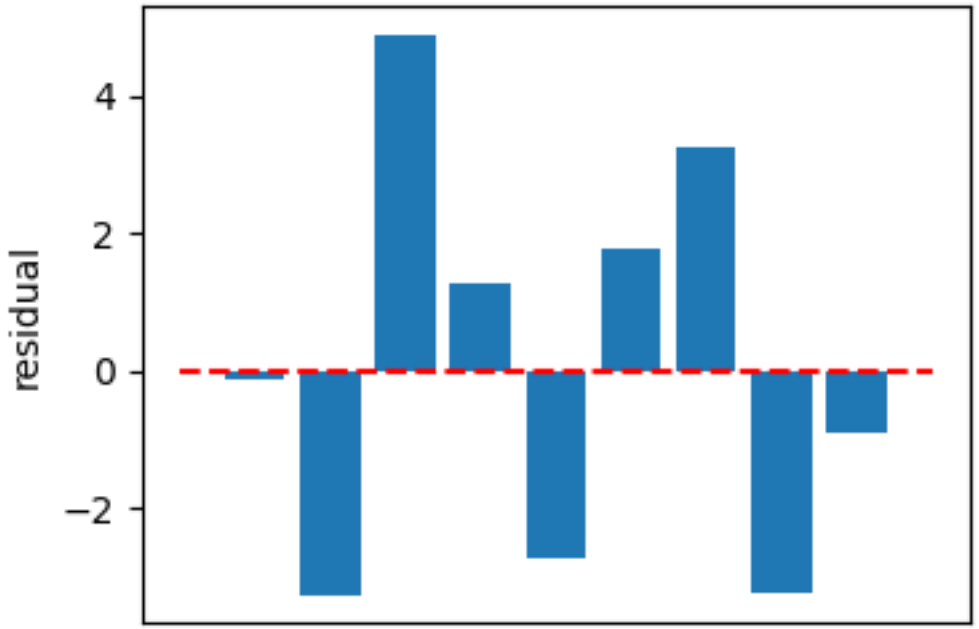
Regression Model Evaluation Metrics

- Regression model:
  - Predict continuous numeric outcomes from input variables.
- Evaluation goal:
  - Evaluate the performance of regression model by measuring the accuracy of predictions.
    - Validate model performance on unseen data.
    - Understand how accurate the prediction is.
    - Compare the performance of different models
- Evaluation metrics:
  - Measure the model performance quantitatively

Regression Model Evaluation Metrics

- For a single instance of data (x_i, y_i), the accuracy of the prediction $\hat{y}_i$ can be measured by the **residual ϵ_i**
  - $\epsilon_i = y_i - \hat{y}_i$
  - Lower residuals are better
  - Measure the accuracy of a single prediction made by the model
  - Cannot reflect the overall accuracy of the regression model

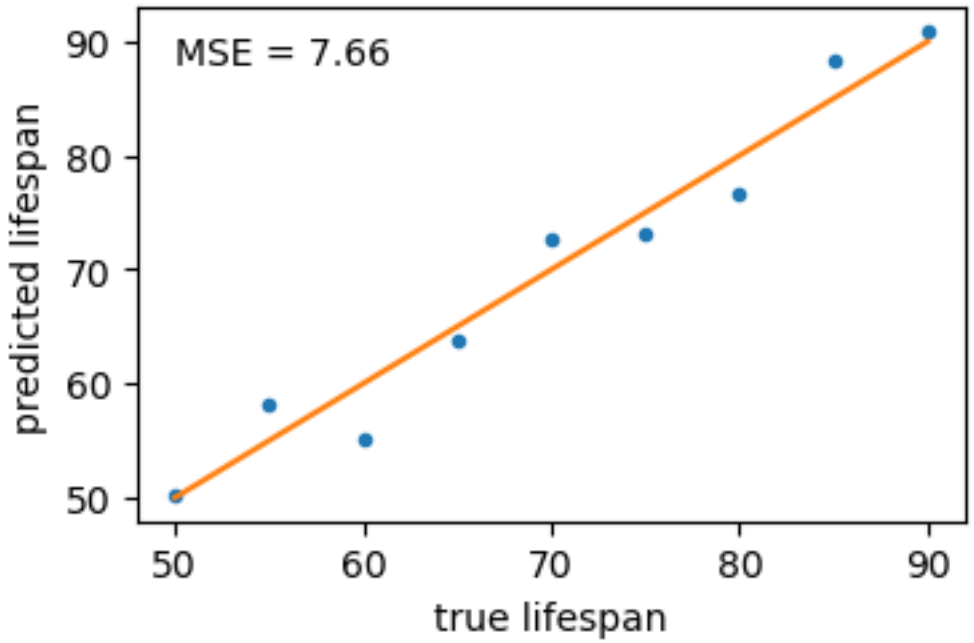
y_i	50	55	60	65	70	75	80	85	90
$\hat{y}_i$	50	58	55	64	73	73	77	88	91
ϵ_i	0	-3	5	1	-3	2	3	-3	-1



Regression Model Evaluation Metrics

- **Mean Squared Error (MSE)**
  - $\text{MSE} = \frac{1}{n}\sum_{i=0}^{n}\epsilon_i^2 = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$
  - Average of the squares of the errors
  - Lower values are better
  - MSE is heavily affected by outliers since the errors are squared before they are averaged

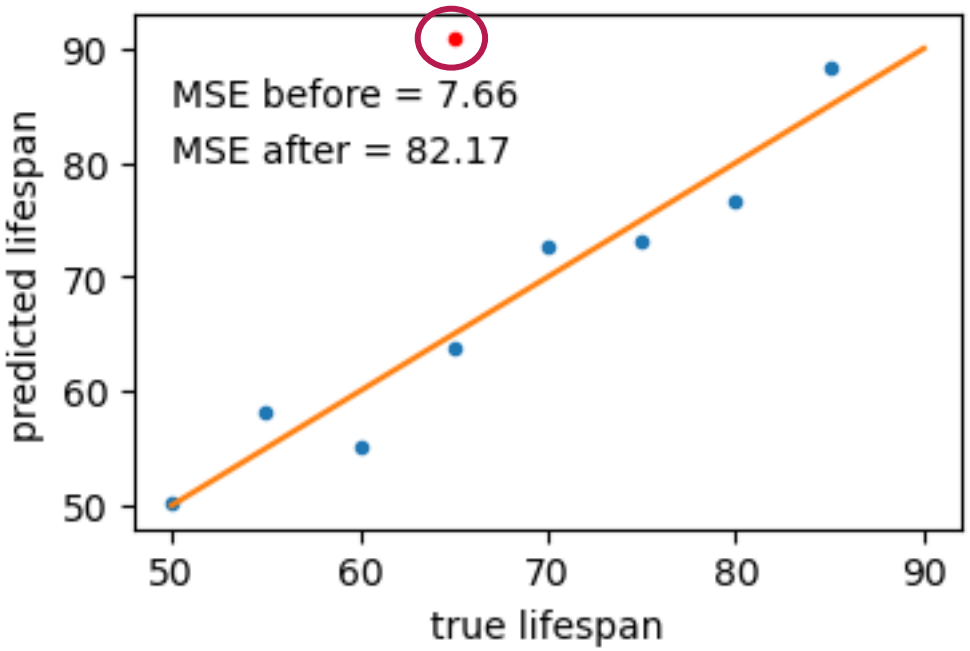
y_i	50	55	60	65	70	75	80	85	90
$\hat{y}_i$	50	58	55	64	73	73	77	88	91
ϵ_i	0	-3	5	1	-3	2	3	-3	-1



Regression Model Evaluation Metrics

- **Mean Squared Error (MSE)**
  - $\text{MSE} = \frac{1}{n}\sum_{i=0}^{n}\epsilon_i^2 = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$
  - Average of the squares of the errors
  - Lower values are better
  - MSE is heavily affected by outliers since the errors are squared before they are averaged
    - There is an outlier: y_9 should be 90, incorrectly recorded as 65.
    - The MSE increases from 7.66 to 82.17
    - **Eliminate outliers before using MSE**

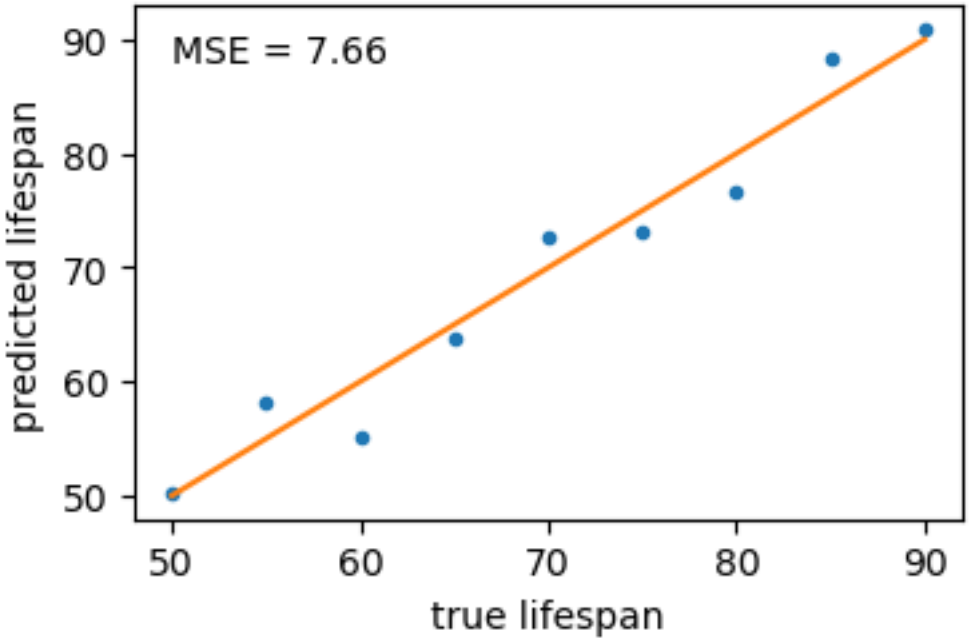
y_i	50	55	60	65	70	75	80	85	***65***
$\hat{y}_i$	50	58	55	64	73	73	77	88	91
ϵ_i	0	-3	5	1	-3	2	3	-3	-1



Regression Model Evaluation Metrics

- **Mean Squared Error (MSE)**
  - $\text{MSE} = \frac{1}{n}\sum_{i=0}^{n}\epsilon_i^2 = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$
  - Average of the squares of the errors
  - Lower values are better
  - MSE is heavily affected by outliers since the errors are squared before they are averaged
  - The unit of MSE is the squared unit of the target variable.
    - Lifespan (year), MSE of lifespan (year^2)

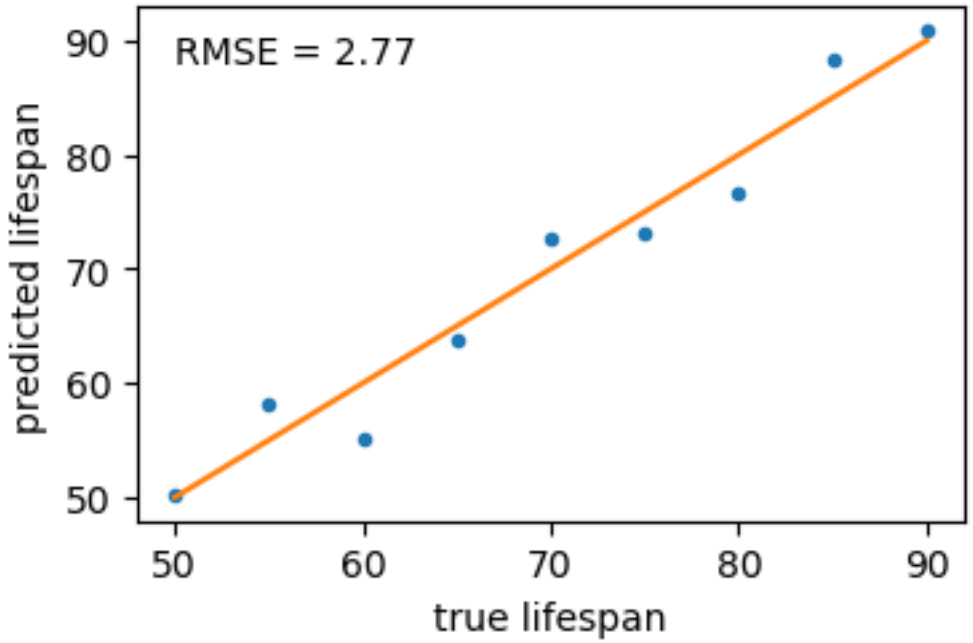
y_i	50	55	60	65	70	75	80	85	90
$\hat{y}_i$	50	58	55	64	73	73	77	88	91
ϵ_i	0	-3	5	1	-3	2	3	-3	-1



Regression Model Evaluation Metrics

- **Root Mean Squared Error (RMSE)**
  - $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$
  - Square root of the MSE
  - Lower values are better
  - RMSE provides error in the units of the target variable.
    - You could say the difference between the true and predicted lifespan is around 2.77 years.

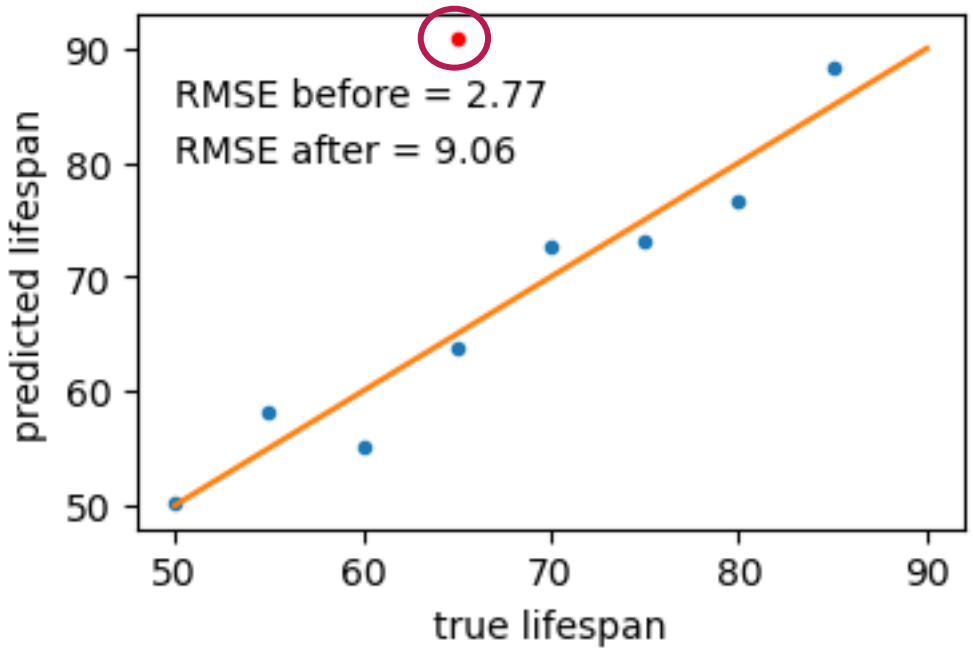
y_i	50	55	60	65	70	75	80	85	90
$\hat{y}_i$	50	58	55	64	73	73	77	88	91
ϵ_i	0	-3	5	1	-3	2	3	-3	-1



Regression Model Evaluation Metrics

- **Root Mean Squared Error (RMSE)**
  - $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$
  - Square root of the MSE
  - Lower values are better
  - RMSE provides error in the units of the target variable.
  - RMSE is still affected by outliers.
    - MSE increased around 10 times.
    - RMSE increased around 3 times.

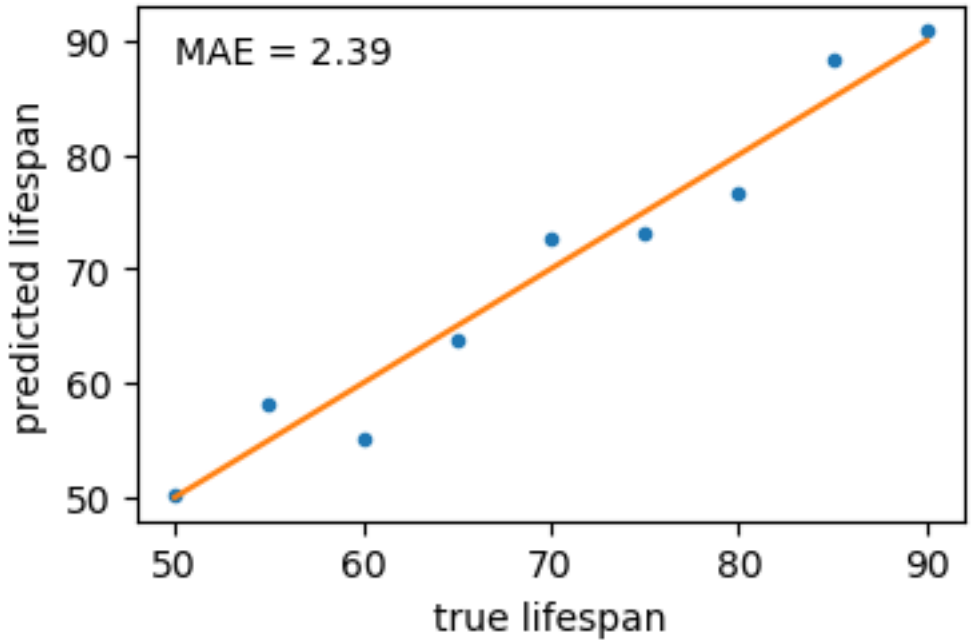
y_i	50	55	60	65	70	75	80	85	***65***
$\hat{y}_i$	50	58	55	64	73	73	77	88	91
ϵ_i	0	-3	5	1	-3	2	3	-3	-1



Regression Model Evaluation Metrics

- **Mean Absolute Error (MAE)**
  - $\text{MAE} = \frac{1}{n}\sum_{i=0}^{n}|\epsilon_i| = \frac{1}{n}\sum_{i=0}^{n}|y_i - \hat{y}_i|$
  - Average of the absolute errors
  - Lower values are better
  - MAE also provides error in the units of the target variable.
    - You could say the difference between the true and predicted lifespan is around 2.39 years.

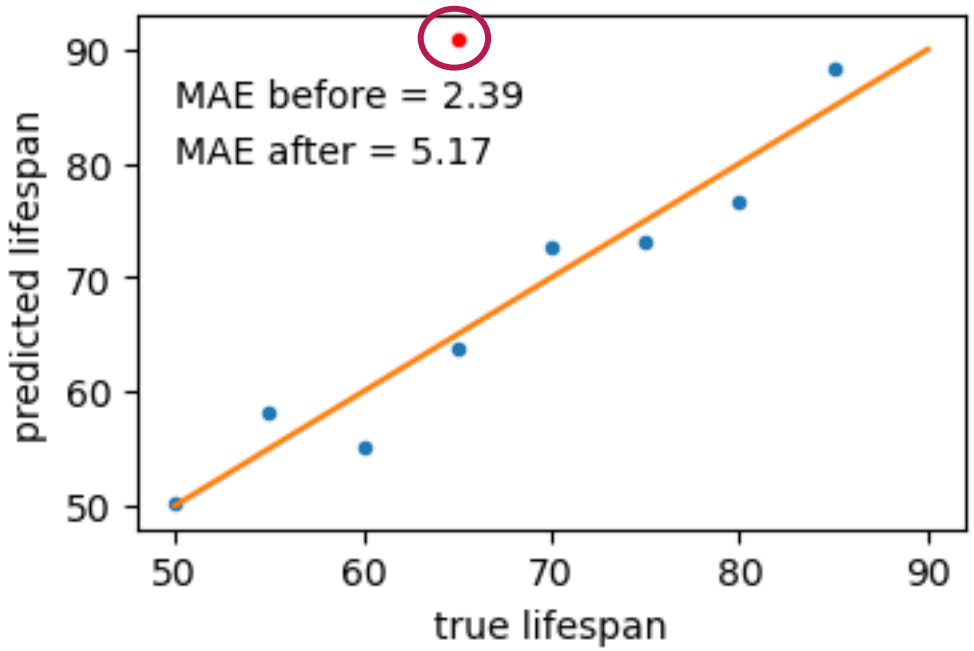
y_i	50	55	60	65	70	75	80	85	90
$\hat{y}_i$	50	58	55	64	73	73	77	88	91
ϵ_i	0	-3	5	1	-3	2	3	-3	-1



Regression Model Evaluation Metrics

- **Mean Absolute Error (MAE)**
  - $\text{MAE} = \frac{1}{n}\sum_{i=0}^{n}|\epsilon_i| = \frac{1}{n}\sum_{i=0}^{n}|y_i - \hat{y}_i|$
  - Average of the absolute errors
  - Lower values are better
  - **MAE is less affected by outliers.**
    - MSE increased around 10 times.
    - RMSE increased around 3 times.
    - MAE increased around 2 times.
    - When there are many outliers in your dataset, or you didn't eliminate outliers, you may use MAE instead of RMSE or MSE.

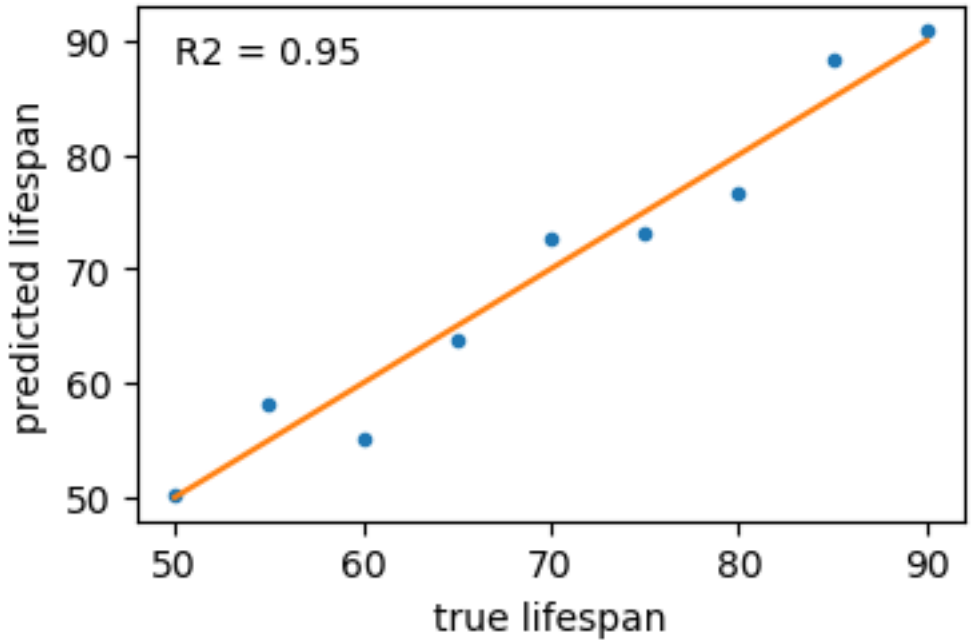
y_i	50	55	60	65	70	75	80	85	***65***
$\hat{y}_i$	50	58	55	64	73	73	77	88	91
ϵ_i	0	-3	5	1	-3	2	3	-3	-1



Regression Model Evaluation Metrics

- **Coefficient of Determination (R^2)**
  - $R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2}$
  - $\bar{y} = \frac{1}{n}\sum_{i=0}^{n} y_i$
  - Range from 0 to 1
  - Values closer to 1 indicate a better fit.

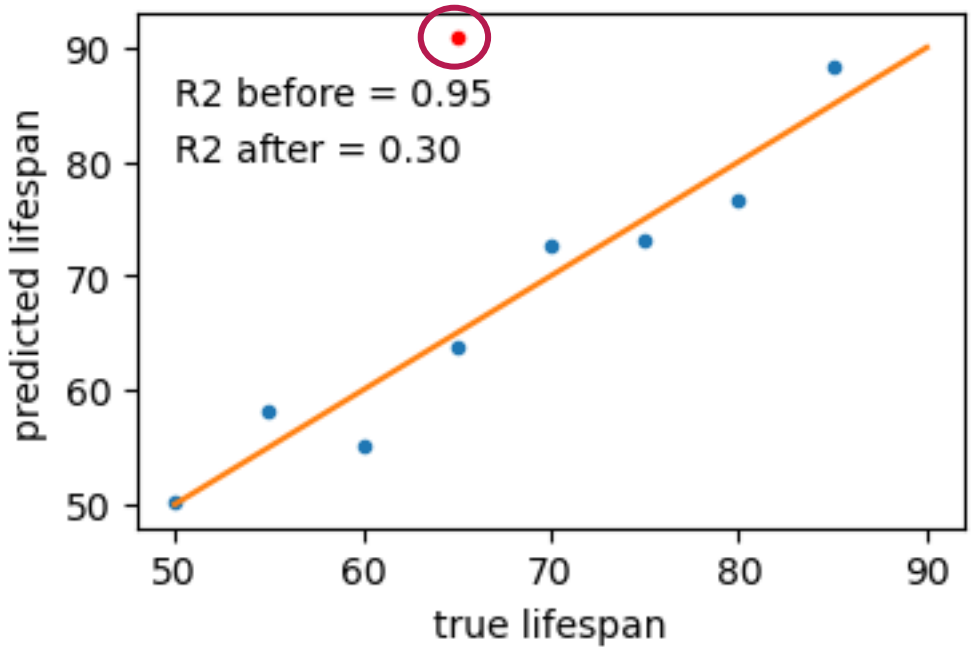
y_i	50	55	60	65	70	75	80	85	90
$\hat{y}_i$	50	58	55	64	73	73	77	88	91
ϵ_i	0	-3	5	1	-3	2	3	-3	-1



Regression Model Evaluation Metrics

- **Coefficient of Determination (R^2)**
  - $R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2}$
  - $\bar{y} = \frac{1}{n}\sum_{i=0}^{n} y_i$
  - Range from 0 to 1
  - Values closer to 1 indicate a better fit.
  - R^2 is also affected by outliers.

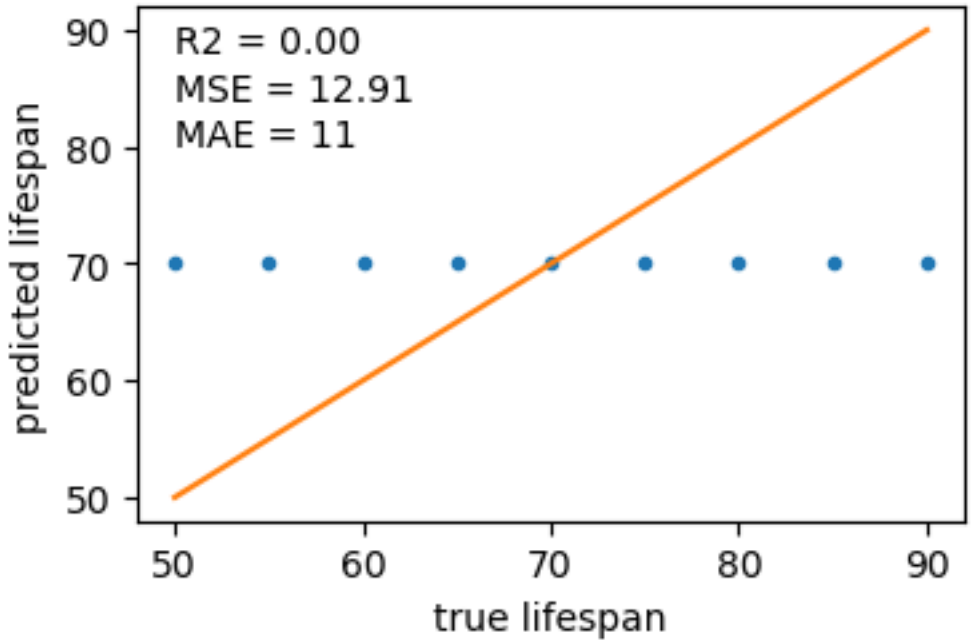
y_i	50	55	60	65	70	75	80	85	***65***
$\hat{y}_i$	50	58	55	64	73	73	77	88	91
ϵ_i	0	-3	5	1	-3	2	3	-3	-1



Regression Model Evaluation Metrics

- **Coefficient of Determination (R^2)**
  - $R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2}$
  - $\bar{y} = \frac{1}{n}\sum_{i=0}^{n} y_i$
  - Range from 0 to 1
  - Values closer to 1 indicate a better fit.
  - When $R^2 = 0$:
    - The regression model learned nothing; it just return the value of $\bar{y}$ instead of giving a meaningful prediction.
    - MSE and MAE may still be acceptable.

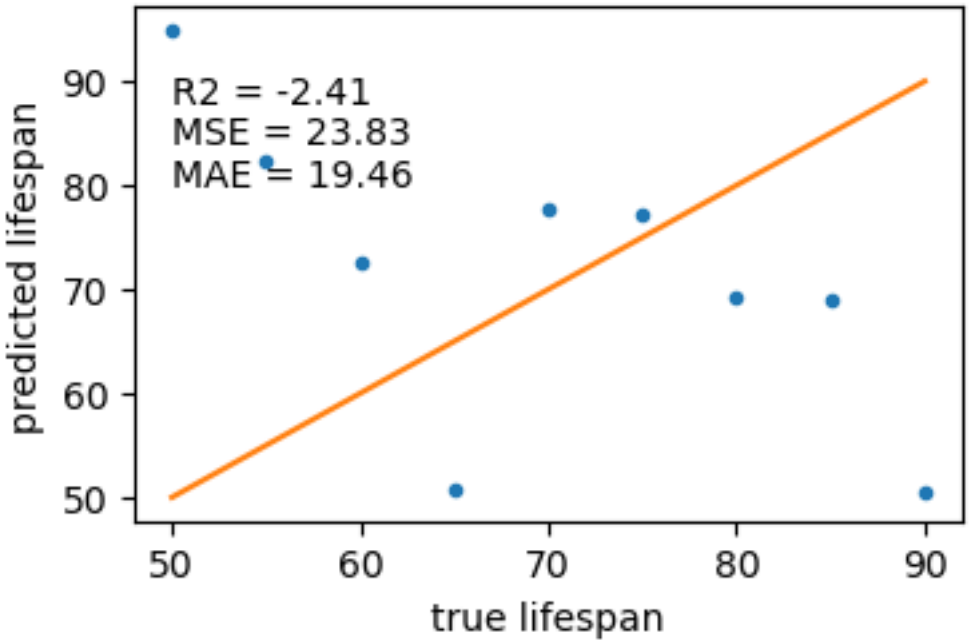
y_i	50	55	60	65	70	75	80	85	90
$\hat{y}_i$	***70***	***70***	***70***	***70***	***70***	***70***	***70***	***70***	***70***
ϵ_i	-20	-15	-10	-5	0	5	10	15	20



Regression Model Evaluation Metrics

- **Coefficient of Determination (R^2)**
  - $R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2}$
  - $\bar{y} = \frac{1}{n}\sum_{i=0}^{n} y_i$
  - Range from 0 to 1
  - Values closer to 1 indicate a better fit.
  - When $R^2 < 0$:
    - When the residuals take really large values
    - The regression model learned wrong patterns
    - Less intuitional from MAE and MSE

y_i	50	55	60	65	70	75	80	85	90
$\hat{y}_i$	*95*	*82*	*73*	*51*	*78*	*77*	*69*	*69*	*51*
ϵ_i	*-45*	*-27*	*-13*	*14*	*-8*	*-2*	*11*	*16*	*39*



Hands-on Exercises

- SLR with ordinary least squares approach
- SLR with sklearn.linear_model.LinearRegression