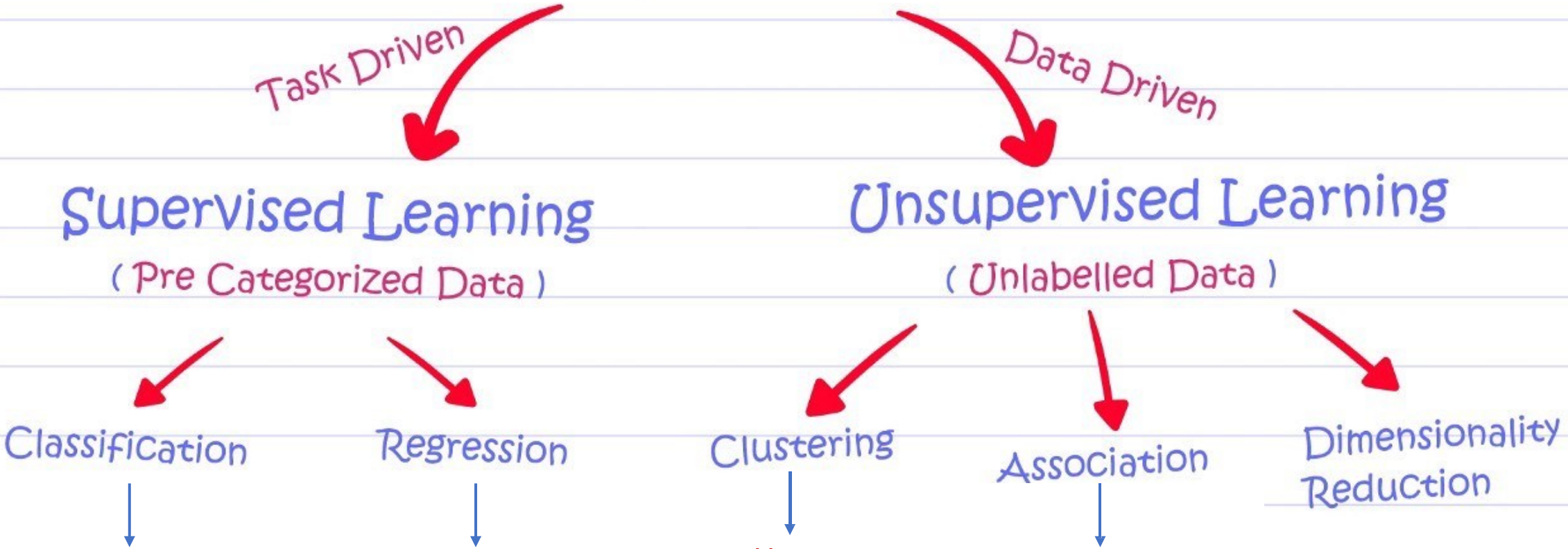
Regression

Summer 2024

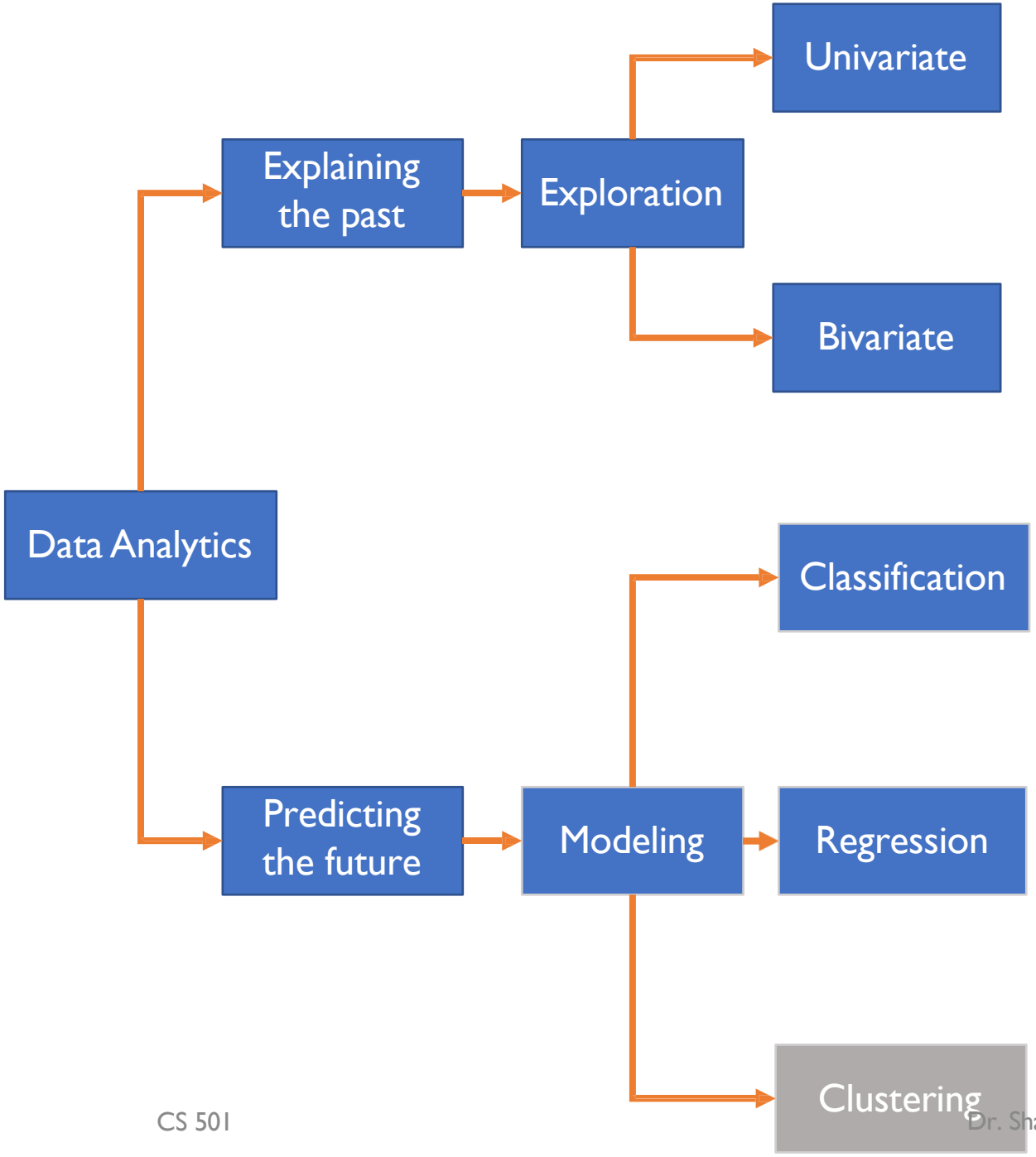
Classical Machine Learning

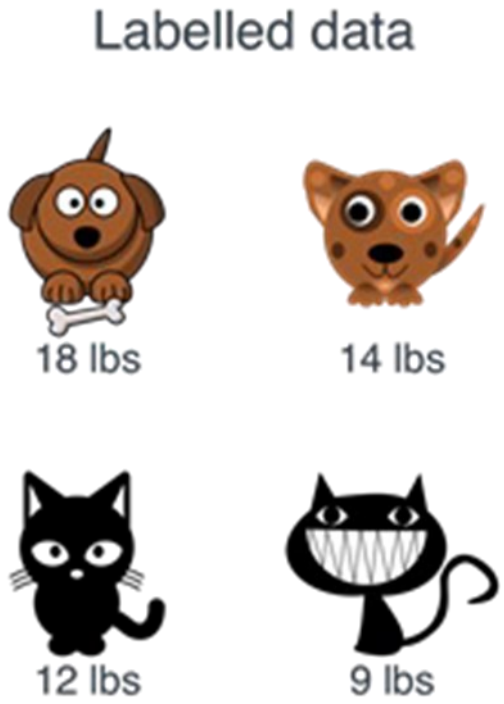


Output:

- A category
- A real-value
- patterns within a group of uncategorized data
- identify associations between different data objects
- identify wider dependencies

Objective: Predictive analysis — Pattern recognition


Univariate
Explaining the past
Exploration
Bivariate
Data Analytics
Classification
Predicting the future
Modeling
Regression
Clustering


Labelled data
18 lbs
14 lbs
12 lbs
9 lbs

What is Simple Linear Regression?

- A statistical method used to model the relationship between two variables by fitting a linear equation to the observed data.

- The outcome is a prediction of the value of a variable based on the value of another variable.

- The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable (or explanatory or predictor).

Linear regression: Formula

- The formula for simple linear regression is:

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

where

- y is the dependent variable
- x is the independent variable
- β_0 is the intercept (the value of y when x is zero)
- β_1 is the slope (the change in y per unit change in x)
- ε is the error term (the amount by which the actual value of y differs from the predicted value based on the model)

Example

- predicting a person's weight (dependent variable) based on their height (independent variable).

- In this case, the formula for simple linear regression would be:

$$Y = a + bX$$

Where:

- Y is the predicted weight
- a is the intercept or constant term
- b is the slope coefficient
- X is the height

Example

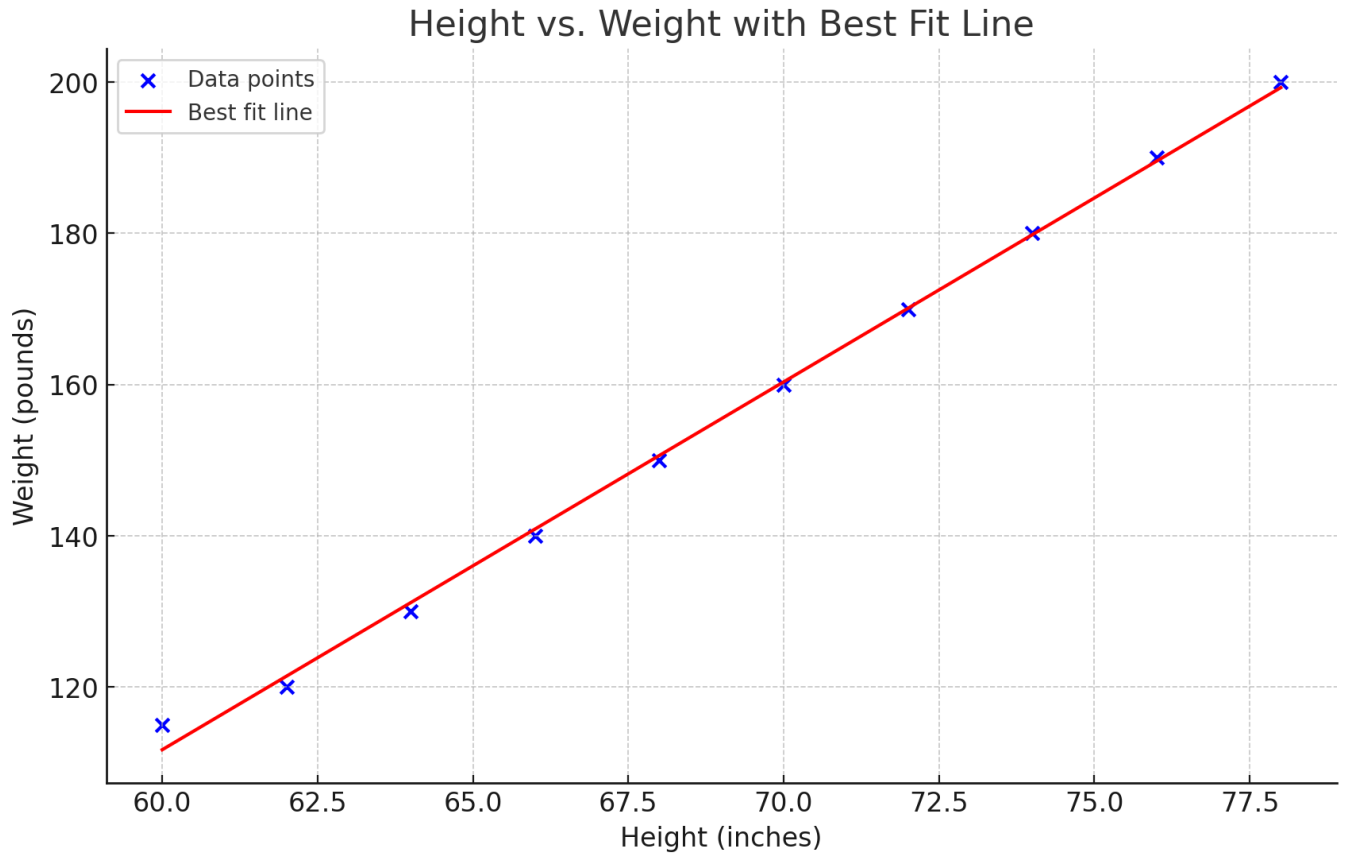
Person	Height (X)	Weight (Y)
1	60	115
2	62	120
3	64	130
4	66	140
5	68	150
6	70	160
7	72	170
8	74	180
9	76	190
10	78	?

Compute $\sum X$, $\sum Y$, $\sum X^2$ and $\sum XY$.
Use the formulas to determine a and b.
Formulate the regression equation $Y=a+bX$

$$a = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum x^2) - (\sum x)^2}$$

Example



Scatter plot of height versus weight with the best fit line (regression line) displayed in red. This line represents the predicted weights based on height using the linear regression equation.

Slope indicates how much the target variable changes with the increase in independent or dependent variables.

Linear regression: Assumptions

1. Linearity: The relationship between the independent and dependent variable is linear.

2. Independence: The observations are independent of each other/no autocorrelation.

3. Homoscedasticity: The variance of the errors is the same for all values of the independent variable.

Homoscedasticity: more

- Heteroscedasticity refers to the phenomenon where the variance of the residuals in a regression model is not constant across all values of the independent variable(s).

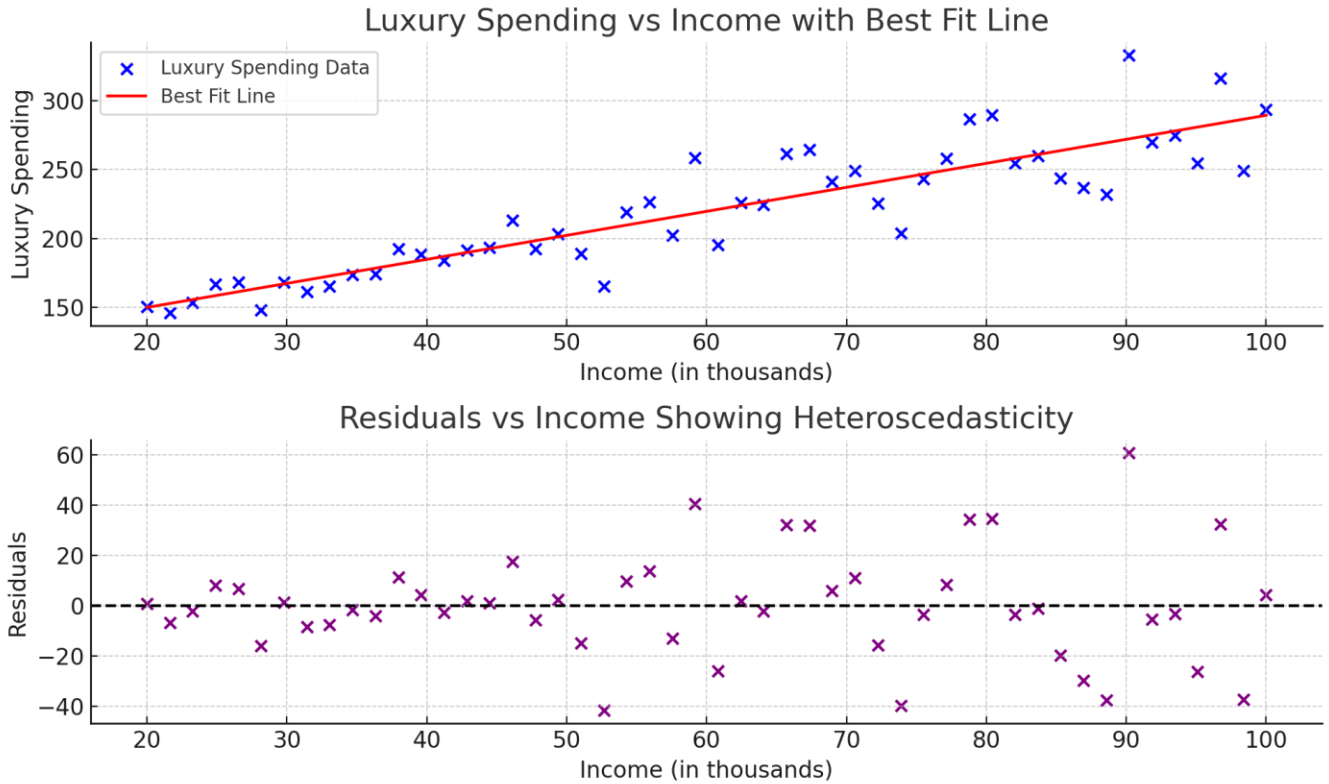
- In other words, the spread of the residuals* changes as the value of the independent variable(s) changes.

- *residuals: difference between the predicted value and the actual value of the dependent variable.

Heteroscedasticity Example

Income (independent variable) and amount spent on luxury goods (dependent variable).

For low-income levels, residuals are close to zero, showing small variance. For high-income levels, residuals vary greatly above and below zero, indicating increasing variance with income



Steps to Identify Heteroscedasticity

1. **Create a Regression Model:**
2. **Analyze Residuals**: After fitting the model, calculate residuals
3. **Plot Residuals vs. Fitted Values**: Plot the residuals against the fitted values (predicted value). In a situation with **homoscedasticity** (constant variance), the residuals would form a "cloud" with roughly equal spread across all levels of income.
4. **Look for Patterns**:
    1. In this example, at lower income levels, the residuals are tightly clustered around the regression line (indicating lower variance in spending).
    2. However, as income increases, the residuals start to spread out, forming a "funnel shape" where residuals vary more widely. This is **heteroscedasticity**—the variance of the residuals increases with income.

Homoscedasticity: Example

- let's consider a simple linear regression model where we are trying to predict a person's salary based on their age.

- If there is no heteroscedasticity in this model, it means that the variance of the residuals (the difference between the predicted salary and the actual salary) changes as the age of the person changes.

- One possible reason for this could be that younger individuals with lower salaries tend to have a higher variance in their salaries compared to older individuals with higher salaries.

Linear regression: Assumptions

4. Normality: The residuals* should be normally distributed.

5. No multicollinearity: There should be no high correlation between the independent variables.

6. Constant variance: Residuals should have constant variance.

*residuals: difference between the predicted value and the actual value of the dependent variable.

Solutions to Multicollinearity

1. **Remove One of the Correlated Variables**: If two variables are highly correlated, consider dropping one from the model.
2. **Combine Variables**: Sometimes creating a composite variable, like combining "house size" and "number of bedrooms" into "total living area," can resolve the issue.
3. **Principal Component Analysis (PCA)**: PCA transforms correlated variables into a smaller set of uncorrelated components, which can then be used in the regression model.

Linear regression: Limitations

1. Linearity assumption: Linear regression assumes a linear relationship between the dependent and independent variables, which may not always be the case.
2. Outliers: Linear regression is sensitive to outliers, which can impact the model's accuracy.
3. Overfitting: Linear regression may overfit the data if there are too many independent variables or the model is too complex.

Linear regression: Limitations

4. Non-normality of errors: If the residuals are not normally distributed, the model may be inaccurate.
5. Extrapolation: Linear regression should not be used to make predictions outside the range of the independent variable values used to fit the model.

Multiple Linear Regression

What is Multiple Linear Regression?

- A statistical method used to model the relationship between a dependent variable and two or more independent variables.

- It is an extension of simple linear regression, where only one independent variable is used to predict the dependent variable.

- In multiple linear regression, a linear equation is used to model the relationship between the dependent variable Y and the independent variables X1, X2, X3, ...Xn.

Multiple Linear Regression: Formula

- The formula for multiple linear regression is:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n + \varepsilon$$

where:

- Y is the dependent variable we are trying to predict
- X1, X2, X3, ..., Xn are the independent variables used to predict Y
- b0 is the intercept or constant term
- b1, b2, ..., bn are the coefficients or slopes of the independent variables
- ε is the error term or residual

Model Evalutation

Evaluating regression models

- We cannot calculate accuracy for a regression model. Why?

- What can be measures is how close the predictions were to the expected values.

- There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are:
  - Mean Squared Error (MSE).
  - Root Mean Squared Error (RMSE).
  - Mean Absolute Error (MAE)

Mean Squared Error (MSE)

- Mean Squared Error, or MSE for short, is a popular error metric for regression problems.

- The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

Example

- Look at regression notebook - cell 1
- A perfect mean squared error value is 0.0, which means that all predictions matched the expected values exactly.
- This is almost never the case, and if it happens, it suggests your predictive modeling problem is trivial.
- A good MSE is relative to your specific dataset.
- It is a good idea to first establish a baseline MSE for your dataset using a naive predictive model, such as predicting the mean target value from the training dataset. A model that achieves an MSE better than the MSE for the naive model has skill.

Root Mean Squared Error or Deviation (RMSE or RMSD)

- The Root Mean Squared Error, or RMSE, is an extension of the mean squared error.

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

- Importantly, the square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted.

- For example, if your target variable has the units "dollars," then the RMSE error score will also have the unit "dollars" and not "squared dollars" like the MSE.

Example

- Look at regression notebook – cell 2

- A perfect RMSE value is 0.0, which means that all predictions matched the expected values exactly.
- This is almost never the case, and if it happens, it suggests your predictive modeling problem is trivial.
- A good RMSE is relative to your specific dataset.
- It is a good idea to first establish a baseline RMSE for your dataset using a naive predictive model, such as predicting the mean target value from the training dataset. A model that achieves an RMSE better than the RMSE for the naive model has skill.

Mean Absolute Error (MAE)

- Mean Absolute Error, or MAE, is a popular metric because, like RMSE, the units of the error score match the units of the target value that is being predicted.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

  - Unlike the RMSE, the changes in MAE are linear and therefore intuitive.
  - That is, MSE and RMSE punish larger errors more than smaller errors, inflating or magnifying the mean error score. This is due to the square of the error value. The MAE does not give more or less weight to different types of errors and instead the scores increase linearly with increases in error.

Example

- Look at regression notebook - cell 3

- A good MAE is relative to your specific dataset.
- It is a good idea to first establish a baseline MAE for your dataset using a naive predictive model, such as predicting the mean target value from the training dataset. A model that achieves a MAE better than the MAE for the naive model has skill.

R-squared or Coefficient of Determination

- It measures the strength of the relationship between your model and the dependent variable.

- This metric represents the part of the variance of the dependent variable explained by the independent variables of the model.

- R-squared ranges between 0 and 1, with higher values indicating a better fit of the model to the data.

Interpreting the coefficient of determination

Coefficient of determination (R^2)	Interpretation
0	The model **does not** predict the outcome.
Between 0 and 1	The model **partially** predicts the outcome.
1	The model **perfectly** predicts the outcome.

Example

A simple linear regression that predicts students' exam scores (dependent variable) from their study time (independent variable) has an R^2 of .71. From this R^2 value, we know that:

- 71% of the variance in students' exam scores is predicted by their study time
- 29% of the variance in student's exam scores is unexplained by the model

Calculating the coefficient of determination

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Where:

- RSS = sum of squared residuals

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

y_i = i^th value of the variable to be predicted

$f(x_i)$ = predicted value of y_i

n = upper limit of summation

- TSS = total sum of squares

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

n = number of observations

y_i = value in a sample

$\bar{y}$ = mean value of a sample

Case Study

Predicting Tips using Linear Regression

Variance in ML refers to the changes in the model when using different portions of the training data set. Simply stated, variance is the variability in the model prediction—how much the ML function can adjust depending on the given data set. Variance comes from highly complex models with a large number of features

In machine learning, variance is the variability in model predictions when different parts of the training data set are used. It measures how much a model's predictions change when using different training sets.

Variance can be caused by complex models with many features. A model with high variance may:

Overfit

Adapt too closely to the training data, including noise, and perform well on the training data but poorly on new data

Reflect random noise

Instead of the target function, the model may reflect random noise in the training data set

A model with low variance means that sampled data is close to where the model predicted it would be. A model with low variance will:

Generalize well

Have relatively consistent and stable predictions when applied to different subsets of the same dataset or when used on new data

Some examples of machine learning algorithms with low variance include: linear regression, logistics regression, and linear discriminant analysis.

Some examples of machine learning algorithms with high variance include: decision trees, support vector machines, and k-nearest neighbors.