Regression Analysis

Patrick C. Shih

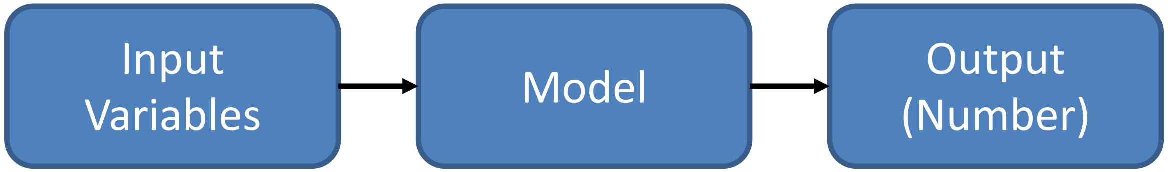
Assistant Professor
Department of Informatics
Indiana University Bloomington

By the end of this lecture, you should be able to

- Understand regression
- Distinguish between regression and classification
- Can build simple regression functions

Regression

- Purpose: predicting with digitized data



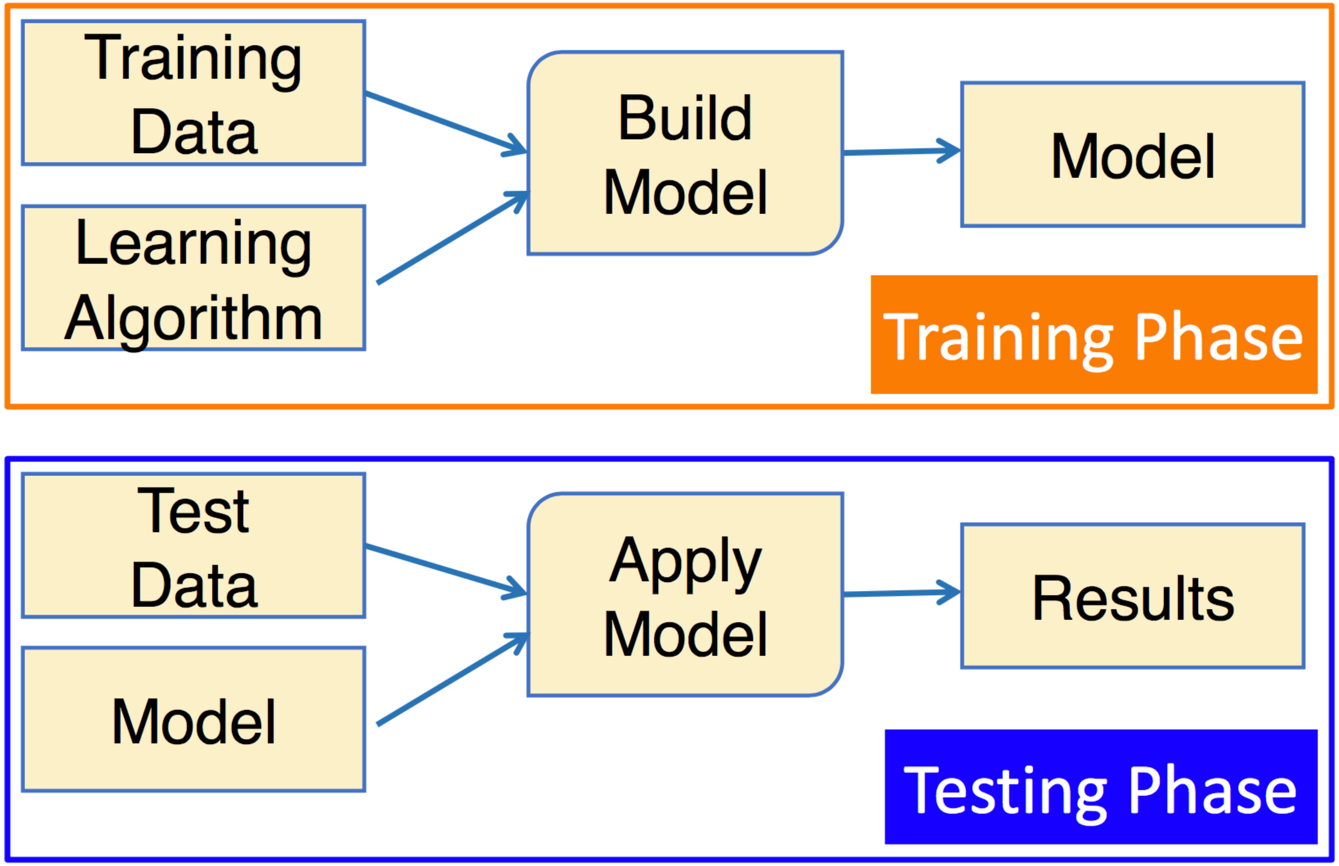
Regression Examples

- Forecast high temperature for next day
- Estimate average house price for a region
- Determine demand for a new product
- Predict power usage

Regression is Supervised Learning

Today's High	Today's Low	Month	Tomorrow's High
79	64	July	81
60	45	October	58
68	49	May	65
57	47	January	54

Training vs Testing



Datasets

Training Data

Model Parameter Adjustments (70~80%)

Validating Data

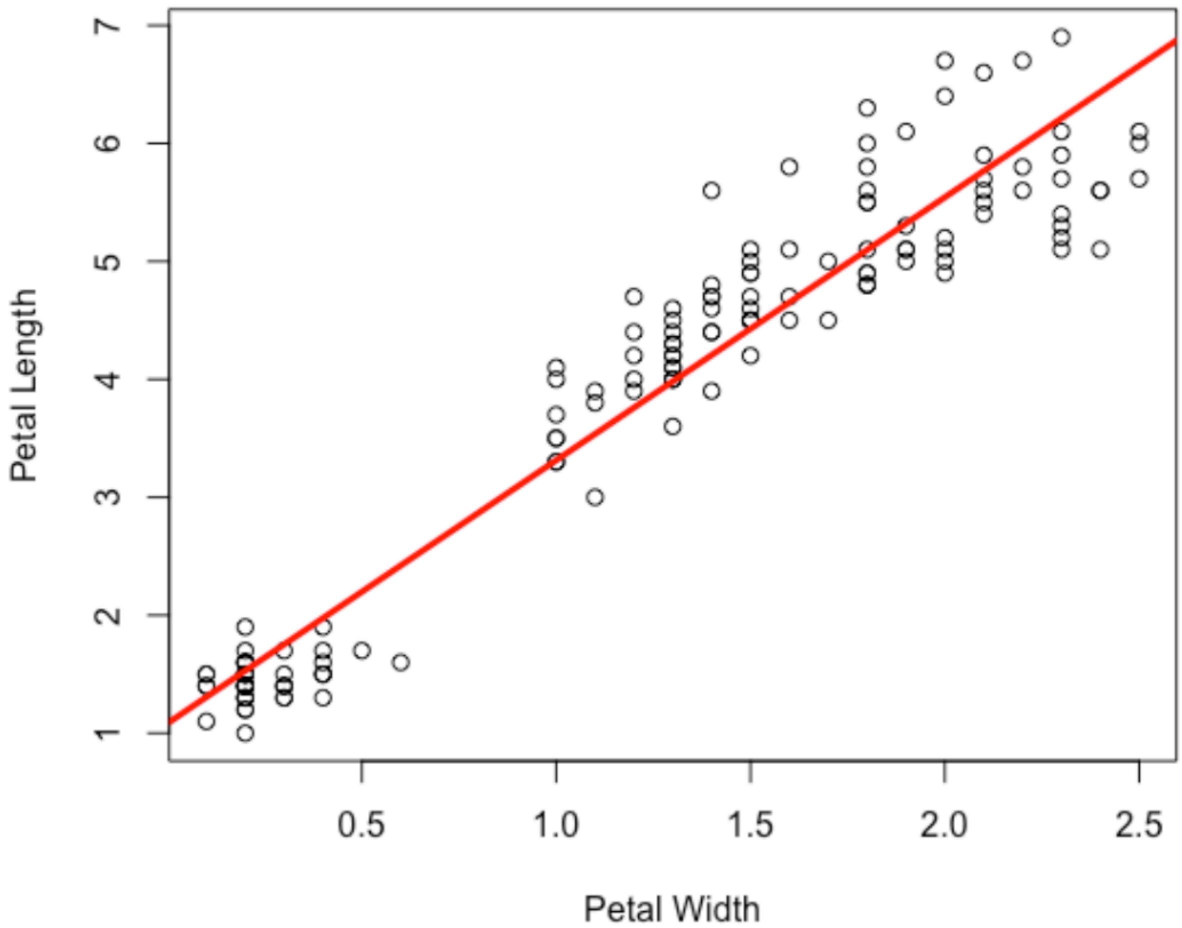
Overfitting to reduce Parameter decision (5~10%)

General performance evaluation

Test Data

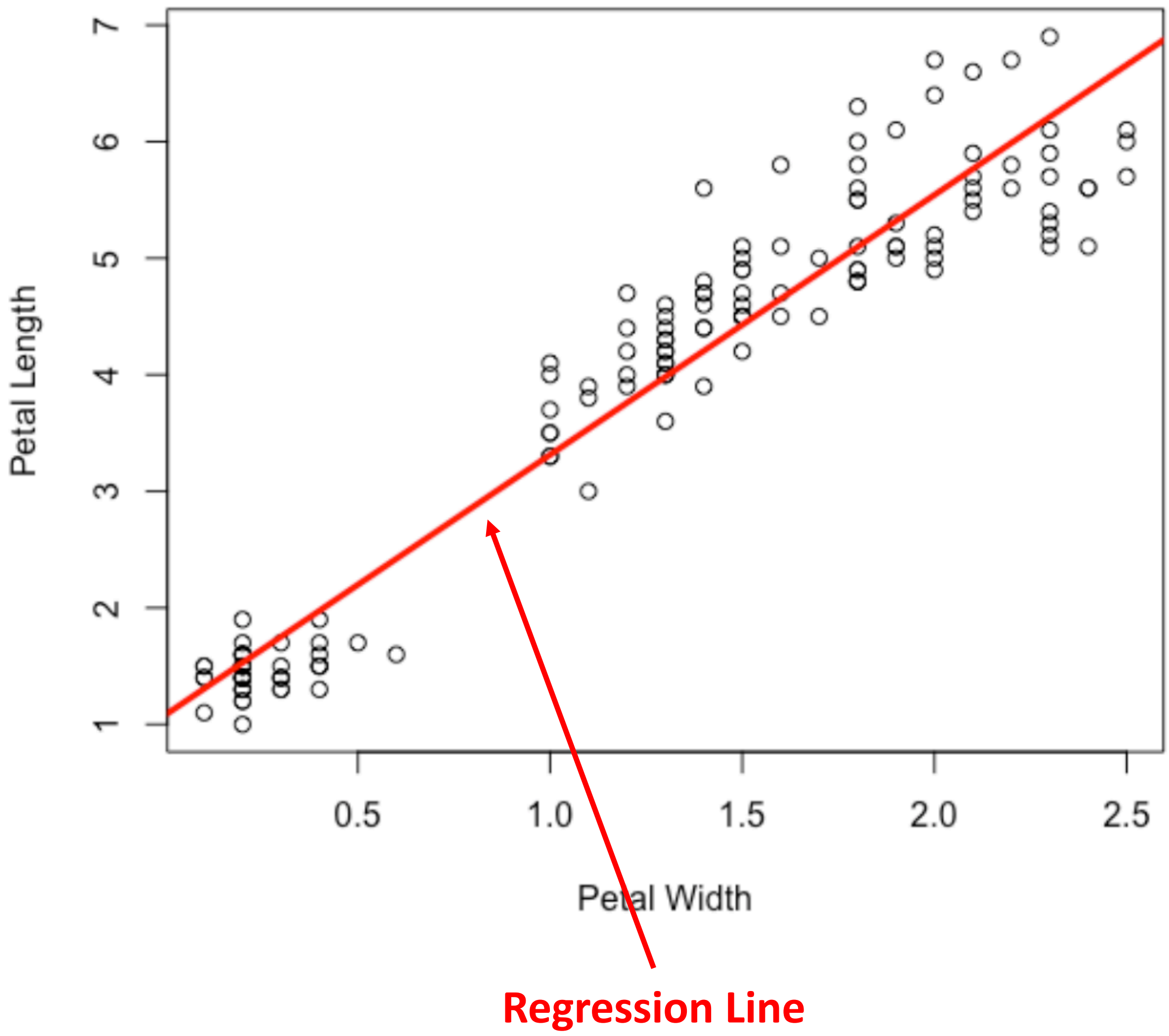
Model on new data Performance evaluation (15~20%)

Linear Regression Model

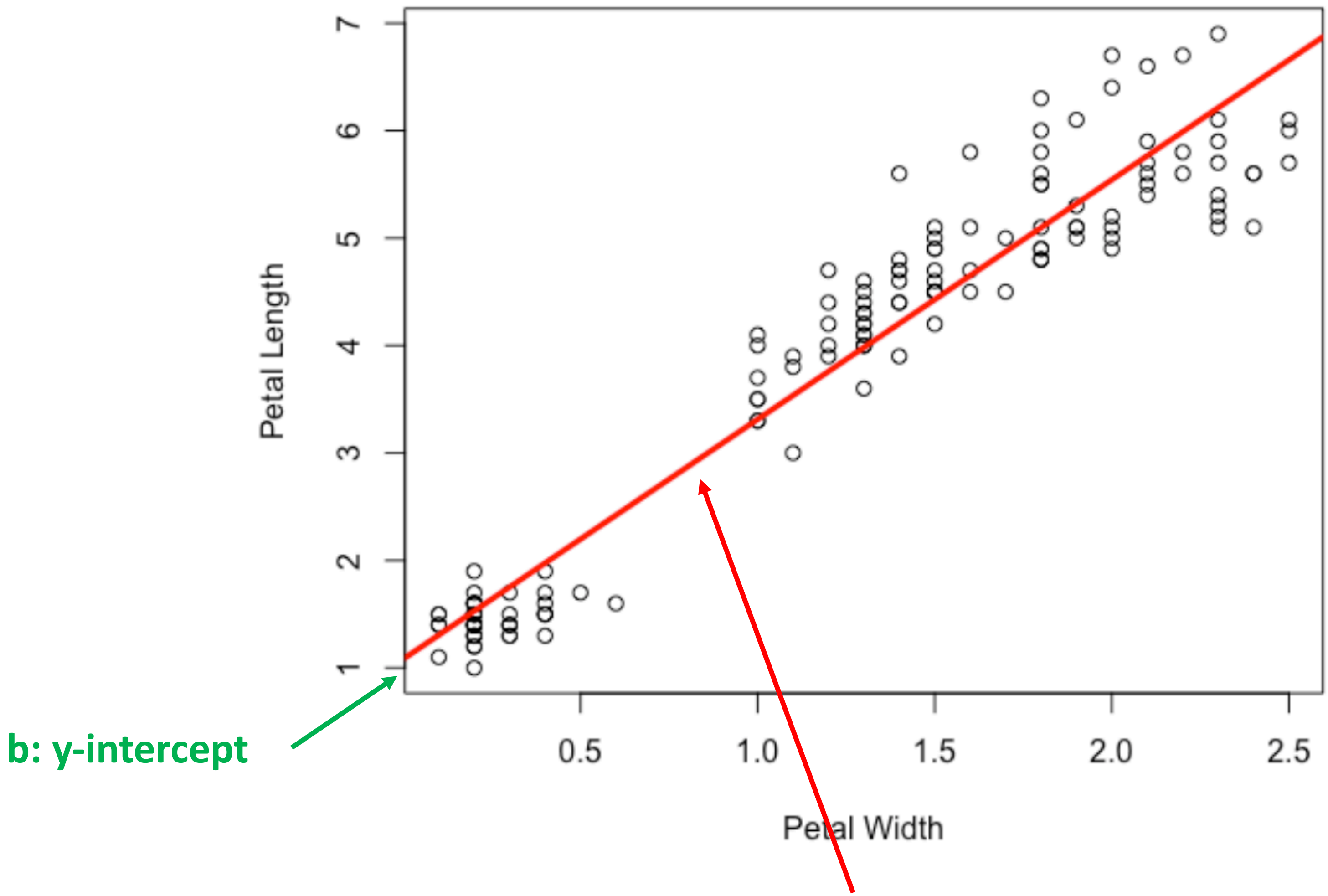


Regression Task: Predict the Petal Length for a given Petal Width.

Linear Regression Model



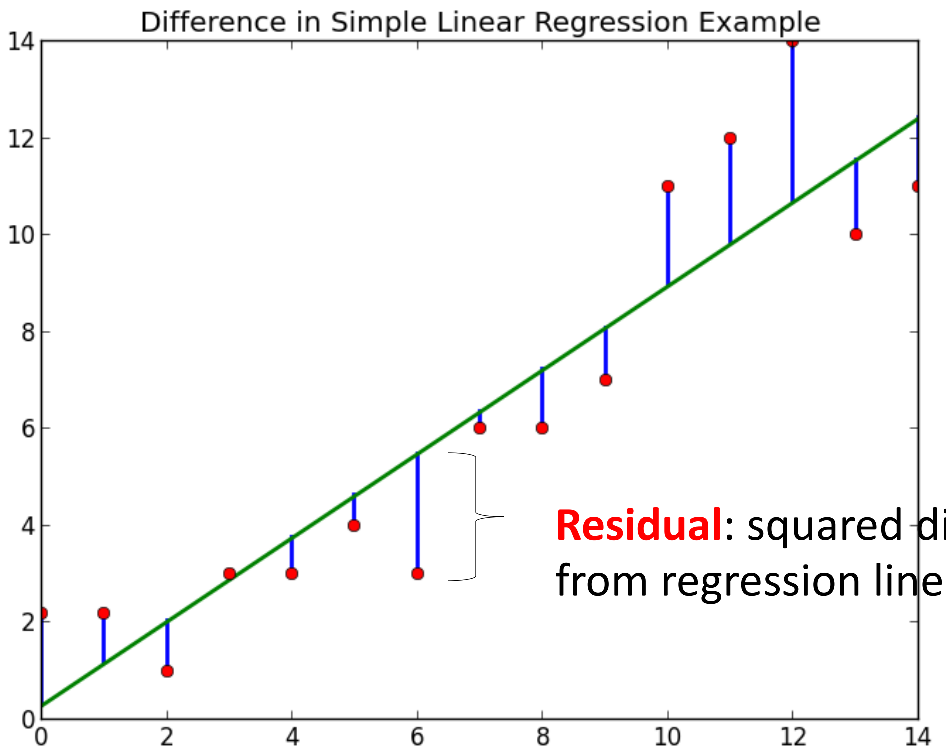
Linear Regression Model



$y = mx + b$

(m and b are model parameters)

Least Square Algorithm



Residual: squared distance from regression line

Goal: find regression line that makes sum of residuals as small as possible

Regression Analysis

$$\boldsymbol{y^{\wedge} = w_0 + w_1 x}$$

- ***x***: explanatory variable
- ***y^***: response or target variable
- ***w0***: y intercept
- ***w1***: variable coefficient

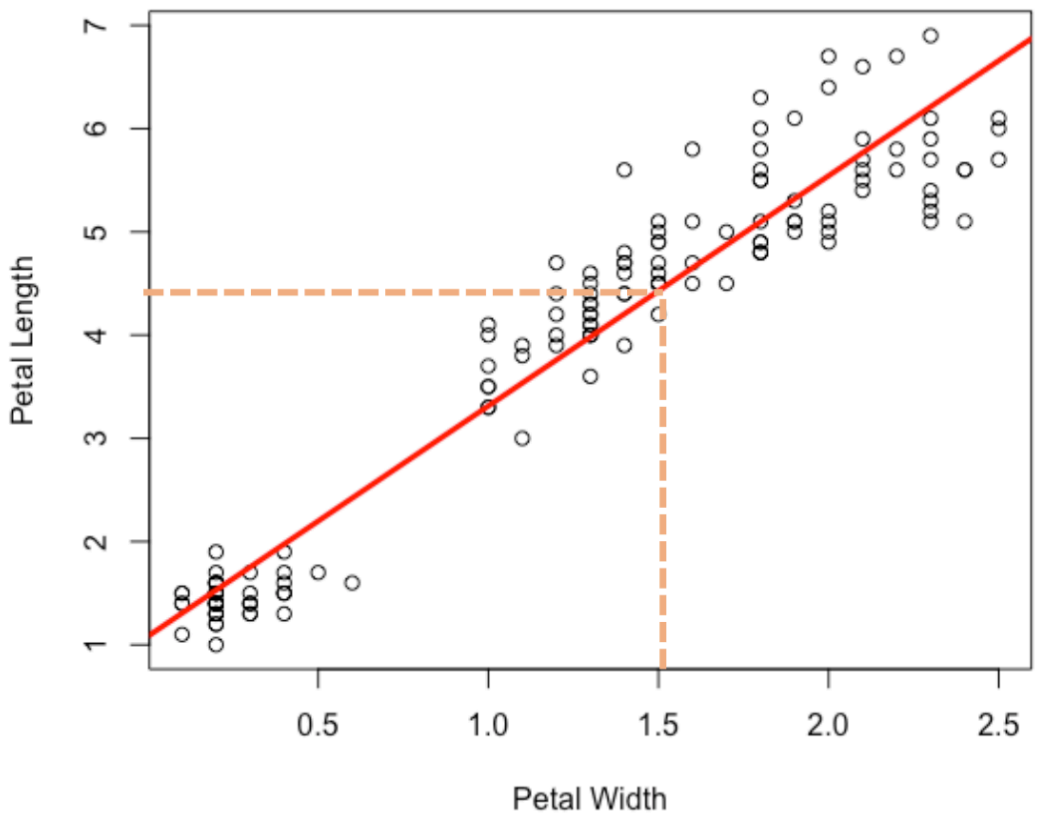
$$\boldsymbol{offset = y^{\wedge} - y}$$

- Offset is the difference between the response (y ^) and the actual response (y)

$$\sum_{i=1}^{n} (y^{\wedge(i)} - y^{(i)})^2$$

- Since the least-squares method squares the offsets for all the data and adds them all up to the minimum, the goal of the regression model is to find a regression model that minimizes the above values.

Linear Regression Model



Applying Model:

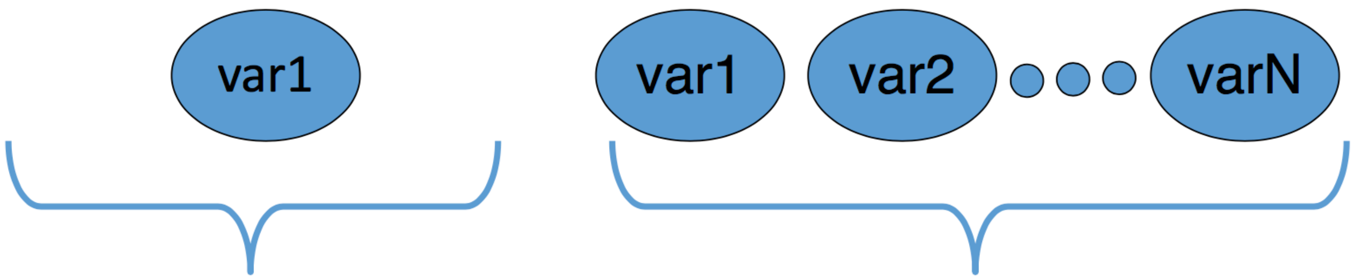
Given petal width = 1.5,

Prediction is petal length = 4.5

Types of Linear Regression

Simple Linear Regression

Multiple Linear Regression



Input has one variable

Input has >1 variables

Evaluating Linear Regression

- *F-statistic*
  - Determine whether the derived regression equations are statistically significant for the entire regression model
- *P-value*
  - Determine if each variable has a significant effect on the dependent variable
- ***R^2 score***
  - Identify the relative proportion of the total change from the change explained by the regression line
  - Determine what percentage of the dependent variable the regression line describes

scikit-learn (or sklearn) library

Simple Linear Regression in sklearn

```
In [1]: import numpy as np
        from sklearn.linear_model import LinearRegression

        x = np.array([[0.0],[1.0],[2.0]])
        y = np.array([1.0, 2.0, 2.9])
```

```
In [2]: lm = LinearRegression()
        lm.fit(x, y)

        print(lm)

        LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

```
In [3]: lm.coef_

Out[3]: array([ 0.95])
```

```
In [4]: lm.intercept_

Out[4]: 1.0166666666666671
```

- Scale
  - Generally means to change the range of the values. The shape of the distribution doesn’t change. Think about how a scale model of a building has the same proportions as the original, just smaller. That’s why we say it is drawn to scale. The range is often set at 0 to 1.
- Standardize
  - Generally means changing the values so that the distribution standard deviation from the mean equals one. It outputs something very close to a normal distribution. Scaling is often implied.
- Normalize
  - Normalizes sample rows, not feature columns, to values between -1 and 1

Coronavirus Data

- https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

Next Class

- Practice two example regression models

Thank you

Patrick C. Shih