$$\text{Machine learning} = \begin{cases} \text{Simple: Predicting sales using tv advertisement} \\ \text{Complex: Spam email detection from inbox} \\ \text{Even more: Detecting frudulent fingerprints} \end{cases}$$

Machine Learning Using R

Professor Dr. Md. Kamrul Hasan

Dept. of ARD, PSTU, Bangladesh

www.pstu.ac.bd/user-profile/129

www.ruenresearch.com

My Background in Statistics

- Undergraduate, BScAg (Hons), BAU
  - Agricultural statistics course
  - Data analysis consultant for small projects
- Masters, MS in Agricultural Extension Education, BAU
  - Data analysis consultant for 2+ PhD and 25+ MS research works
- International Master of Science in Rural Development, IMRD, Belgium, Germany and Slovakia
  - Applied statistics using R programming language
  - Econometrics using R programming language
  - Research methodology using SPSS
- PhD in Ecosystem Management, Environmental and Rural Science, Australia
  - Contributed to 5+ doctoral research projects
  - Worked as a statistics teacher responsible for Machine Learning course using R programming language
  - Published works in high impact (Q1) journals using Machine learning with R
- More importantly, at PSTU as a teacher
  - Conducted several training on research methodology and data analysis using R for university teachers, PhD and MS students
  - Analysed data for numerous research projects using R

Machine Learning Using R

Reference Book:

1. Hands-on machine learning with R (https://bradleyboehmke.github.io/HOML/)

- **Bradley Boehmke and Brandon Greenwell**
- © Taylor & Francis Group, 2020

2. An Introduction to Statistical Learning with Applications in R (https://www.statlearning.com/)

- Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
- © Springer Science+Business Media New York 2013

Machine learning or statistical learning

- Big data – not font size, rather big number of variables and samples
- Predictive task – future is unknown, uncertain
- **Machine learning – statistical tools for understanding data**
- Tools – supervised or unsupervised
- Supervised – known number of classes
- Unsupervised – unknown number of classes
- Usefulness– classification problems

Problem: Regression Classification

Model: Development Validation Prediction Evaluation

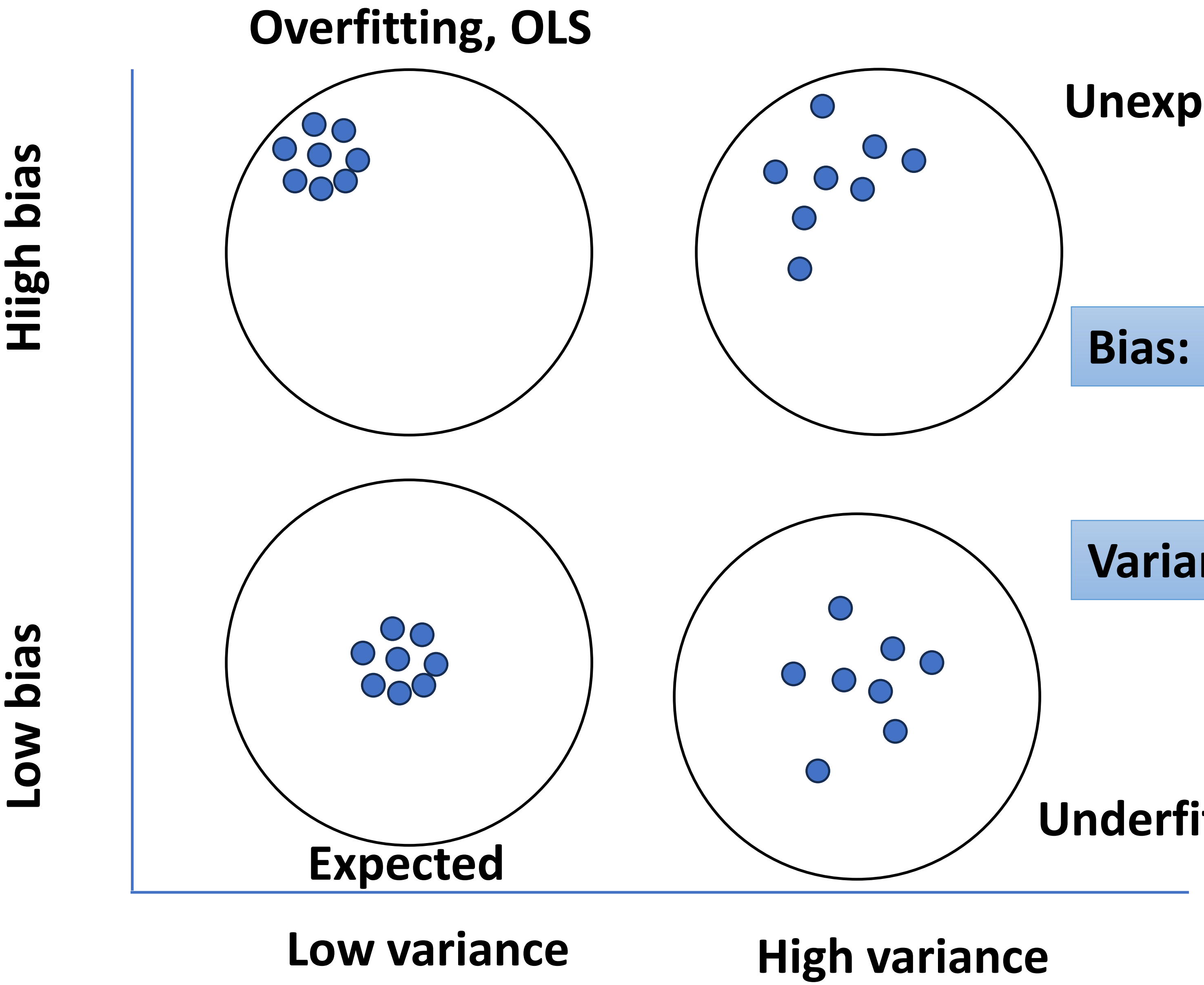
Tools for machine learning

- Linear regression
- Logistic regression
- K-nearest neighbor (KNN)
- Linear discriminant analysis
- Support vector machines
- Random forests
- Artificial neural network

- Cross-validation
- Bootstrap
- Training and test data
- Root Mean Square Error (RMSE)
- Confusion matrix
- Prediction accuracy
- Kappa statistics

Bonus

- Stepwise selection
- Principal components analysis
- K-means clustering
- Hierarchical clustering
- Decomposition of time series



Overfitting, OLS

Unexpected

Hiigh bias

Bias: inability to capture data patterns

Variance: too sensitive to training data

Low bias

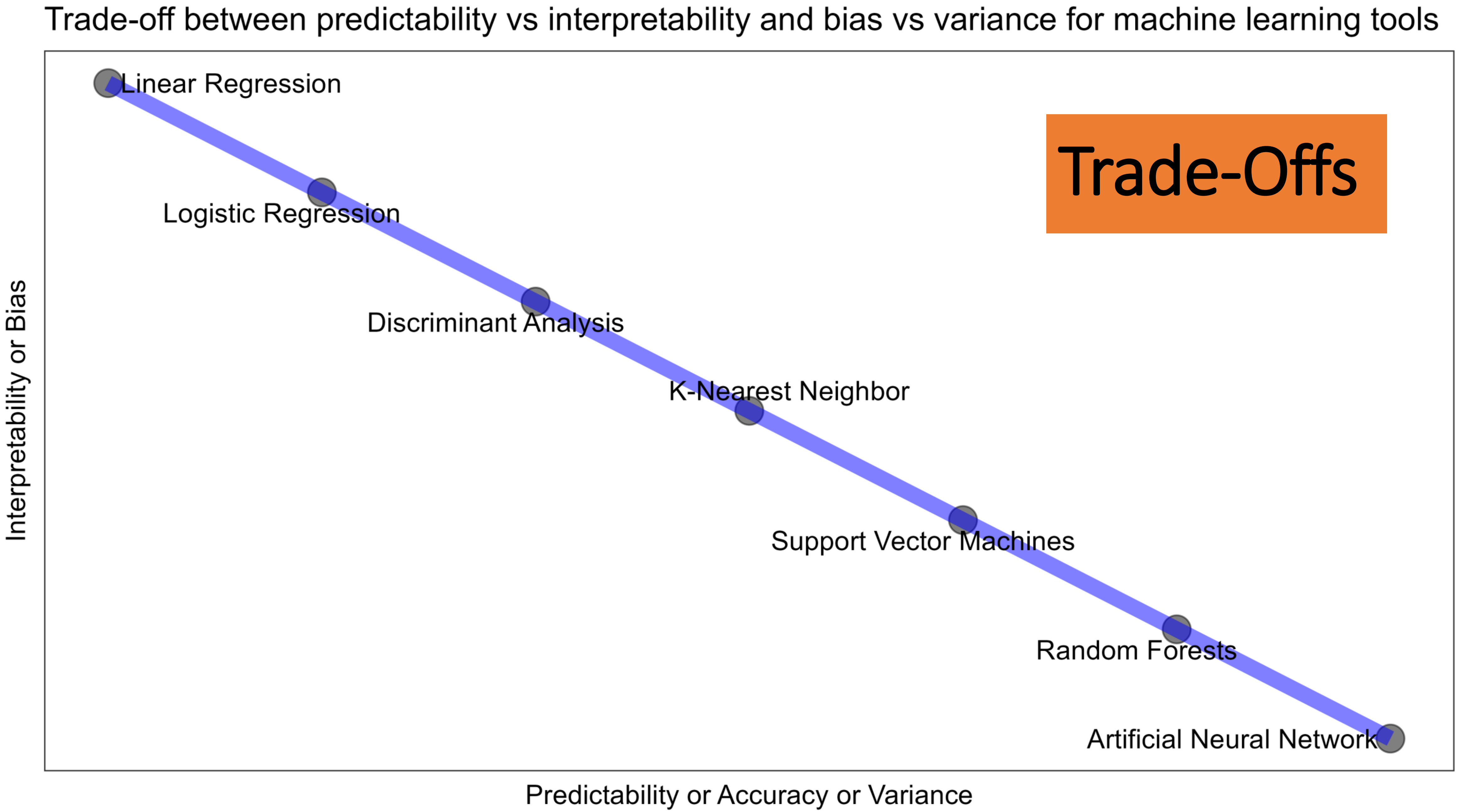
Expected

Underfitting, ANN

Low variance

High variance

Trade-Offs


Trade-off between predictability vs interpretability and bias vs variance for machine learning tools
Linear Regression
Logistic Regression
Discriminant Analysis
K-Nearest Neighbor
Support Vector Machines
Random Forests
Artificial Neural Network
Interpretability or Bias
Predictability or Accuracy or Variance

Supervised vs Unsupervised Learning

- ❑ **Supervised:**
  - ▪ Relatively easy
  - ▪ We know the categories of dependent variables
  - ▪ These categories supervise the analysis
  - ▪ We can analyse emails as spam or not spam
- ❑ **Unsupervised:**
  - ▪ Relatively complex
  - ▪ We need to find out suitable number of categories of the dependent variables
  - ▪ There is no response variable to be predicted
  - ▪ We can create clusters of individuals based on their features
- ❑ **Semi-supervised:**
  - ▪ In between

Regression vs Classification Problem

- **Regression problem**
  - Response variable is quantitative: income, age, height, distance.

- **Classification problem**
  - Response variable is qualitative: yes-no, 0-1, male-female, Benign-Malignent cancer, Brand A-B-C-D.

- Independent or predictor variables can be of either qualitative or quantitative types.

Training Data vs Test Data

- **Training data**
  - ~ part of the dataset which is used to train the model.
  - Number of samples required at least 10 times more than the independent or feature variables
  - If you have sufficient data, 70% of data can be used for training purpose

- **Test data**
  - ~ part of the dataset which is used to validate the model by checking the prediction accuracy
  - Usually, 30% data is kept for validation purpose.

Quality of fit: Mean Squared Error (MSE)

- Estimate the model parameters using training data.
- Predict the dependent variable in the test data (unknown for the model).
- Compare the predicted and observed values of the dependent variable in the test data.
- Calculate the differences as MSE
- $MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{f}(x_i))^2 = Similar\ to\ Variance$
- $RMSE = \sqrt{MSE} = Similar\ to\ Standard\ Deviation$
- Smaller the MSE, better the quality of fit, smaller the bias
- Overfitting: small MSE
- Underfitting: large MSE
- Coefficient of determination: R^2, proportion of variance explained by the model

Measuring Accuracy: Confusion Matrix

- Estimate the model parameters using training data.

- Predict the dependent variable in the test data (unknown for the model).

- Compare with the actual values of the dependent variable in the test data with the predicted values using a confusion matrix.
- AUC: Area under the curve – TPR vs FPR, larger are better

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TP)	False Positives (FP) Type I error
Predicted Negative (0)	False Negatives (FN) Type II error	True Negatives (TN)

Accuracy statistics

	Actually Positive (1)	Actually Negative (0)	Total
Predicted Positive (1)	10	5	15
Predicted Negative (0)	3	15	18
Total	13	20	33

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Type\ I\ error = \frac{FP}{TN + FP}$$

$$Type\ II\ error = \frac{FN}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Accuracy statistics: Kappa statistics

	Actually Positive (1)	Actually Negative (0)	Total
Predicted Positive (1)	10	5	15
Predicted Negative (0)	3	15	18
Total	13	20	33

$$Kappa = \frac{\text{Observed Agreement} - \text{Expected Agreement}}{1 - \text{Expected Agreement}}$$

$$Observed\ Agreement = \frac{TP + TN}{\text{Total}}$$

$$Expected\ Agreement = \frac{\text{Actual P}}{\text{Total}} \times \frac{\text{Predicted P}}{\text{Total}} + \frac{\text{Actual N}}{\text{Total}} \times \frac{\text{Predicted P}}{\text{Total}}$$

- Kappa indicates the probability of agreement obtained not by chance.
- Higher the Kappa, better the agreement.

Interpretation of Kappa Statistics

Kappa (k)	Interpretation
0.81 – 1.00	Almost perfect
0.61 – 0.80	Substantial
0.41 – 0.60	Moderate
0.21 – 0.40	Fair
<0.21	Slight

Resampling: Cross Validation

- Cross validation is a method where k number of subsets of data of the training set is used to build the model. It is called k-fold cross validation.
- Cross validation estimates the prediction error more accurately.
- In 10-fold cross validation, the training data is split into 10 subsets. By rotation, each set is used to develop the model and fit on the remaining set followed by estimating the MSE of predicting the test data.
- The cross validated error is the average of the 10 MSEs.
- More folds, more computation, more accuracy but introduces less bias high variance (overfitting problem).
- Less folds, less computation, less accuracy but more bias and less variance (underfitting problem).
- 10-fold cross validation is somewhat in the middle and can be a better choice to get **lower variance** estimate of the model performance.

Resampling: Bootstrapping

- Bootstrapping Extremely powerful to quantify the model uncertainities
- This method samples (with replacement) the data 100s or 1000s of times to estimate the model coefficients through simulating the same model.
- These 100s orf 1000s of the values for the same coefficients are compared to the actual coefficients to estimate the accuracy of the coefficients.
- Thus, it **reduces bias** in the coefficients, but may introduce higher variance for larger re-sample size compared to the dataset.

Let’s begin …..

Moving the gear from parking to neutral

Some formulas of statistics

- **Frequency distribution**
  - Illiterate 15%, Primary 30%, SSC 50%, Greater that SSC 5%
- **Mean** (arithmetic mean) = total/no. of observation
  - $\bar{x} = \frac{\sum x_i}{n}$
- **Range**: Minimum value, Maximum value [Max - Min]
- **Standard deviation**: $sd, s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$
- **Standard error**: se = sd/√n
- **Coefficient of variation**: cv (%) = (sd/mean)×100
- **Confidence Interval**: $95\%\ \mathrm{CI} = \bar{x} \pm 1.96se$, $99\%\ \mathrm{CI} = \bar{x} \pm 2.58se$

$$Mean, \bar{x} = \frac{\sum x_i}{n}$$

$$Sum\ of\ Squares, SS_x = \sum (x_i - \bar{x})^2$$

$$Variance, var_x = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$Standard\ deviation, s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$Covariance, cov_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$Correlation, r = \frac{cov_{xy}}{s_x s_y}$$

$$Regression, \beta = \frac{cov_{xy}}{var_x} = r\frac{s_y}{s_x}$$

$$Standardized\ regression, \beta^*or\ B = \beta\frac{s_x}{s_y}$$

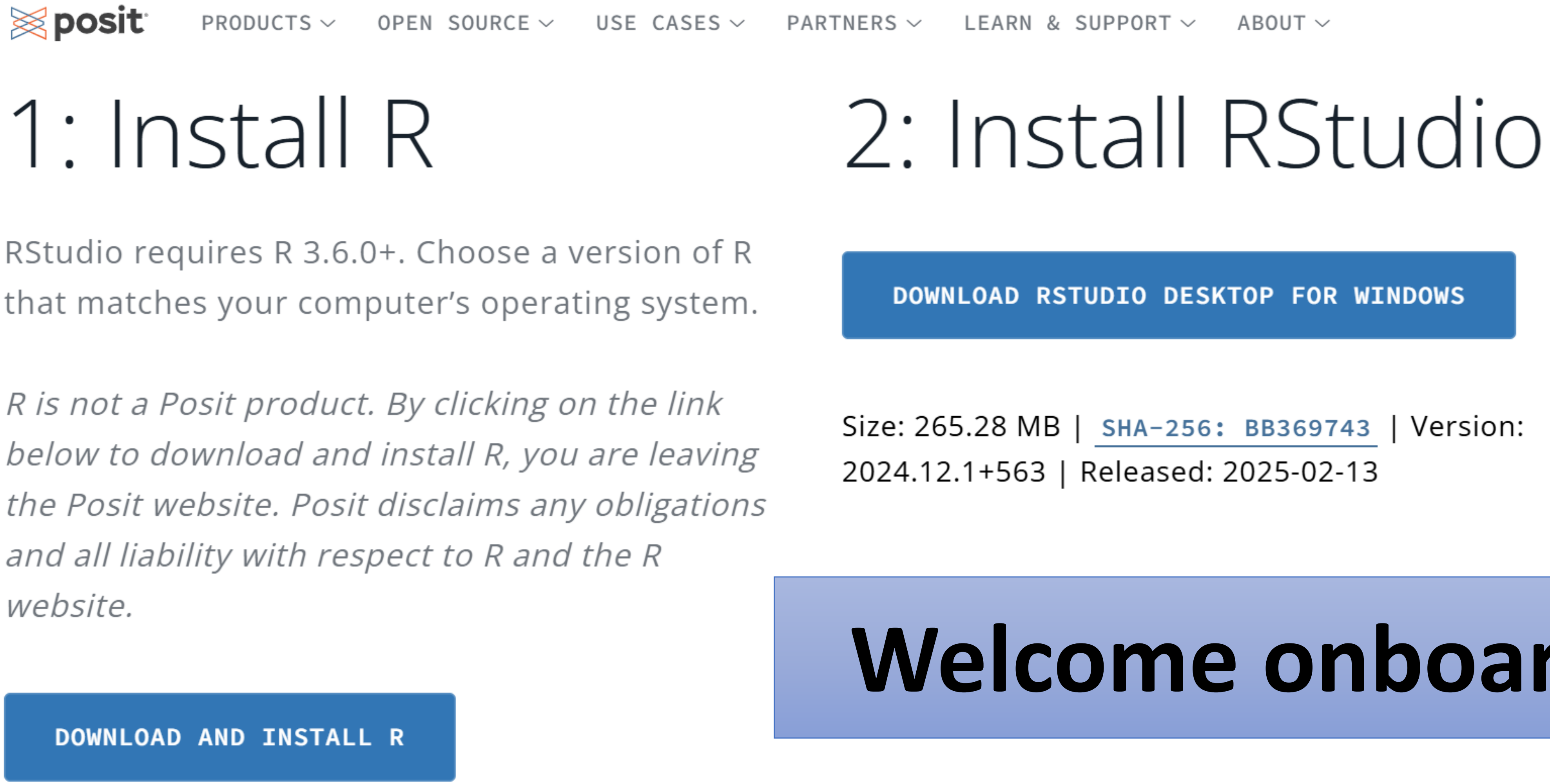
$$Residual\ (error) = Model\ value\ - Actual\ value$$

$$Coefficient\ of\ determination, R^2 = 1 - \frac{Residual\ SS}{Total\ SS}$$

$$Variance\ inflation\ factor, VIF = \frac{1}{1 - R^2}$$

Installation of required software

- Download link <https://posit.co/download/rstudio-desktop/>



1: Install R

RStudio requires R 3.6.0+. Choose a version of R that matches your computer's operating system.

R is not a Posit product. By clicking on the link below to download and install R, you are leaving the Posit website. Posit disclaims any obligations and all liability with respect to R and the R website.

DOWNLOAD AND INSTALL R

2: Install RStudio

DOWNLOAD RSTUDIO DESKTOP FOR WINDOWS

Size: 265.28 MB | SHA-256: BB369743 | Version: 2024.12.1+563 | Released: 2025-02-13

Welcome onboard

Learning paths …

- Creating object
- Object type: scalar, vector/variable, list, matrix, dataframe
- Data type: character, factor, string, numeric, integer, ordered
- Basic functions of R: `length(), sqrt(), mean(), median(), sd(), log(), var()`
- Operators: assignment, arithmetic, logical, modulus/remainder operator
- Custom functions: `se = function (){}, ci = function(){}`
- Conditional statement: `if(){}, ifelse(){}`
- Help function: `help(‘scale’)` or `?scale`
- Advance functions: `apply(), lapply(), sapply()`
- For loop in R: `for (){}`

Learning paths …

- Working with data
  - Create
  - Write
  - Read
  - Slice (select, filter)
  - Manipulate (add variable)
  - Wrangling (pivot_wide, pivot_long)
  - Split (train, test)
  - Sample from the dataset

Learning paths …

- Formula: y ~ x1 + x1 + x……
- Additional functionalities of R: `install.packages(‘pacman’)`
- Loading packages for loading additional functions:
  - `pacman::p_load(tidyverse, jtools, sjstats, caret)`
- Practice all modelling steps using linear regression:
  - `caret funtions: train(), trainControl(), expand.grid(), predict(), confusionMatrix()`
  - `Cross validation and bootstrapping`