Starting with `Rstudio` and What is Statistical Learning?

Section 2.1

Cathy Poliak, Ph.D.
cpoliak@central.uh.edu

Department of Mathematics
University of Houston

Outline

R

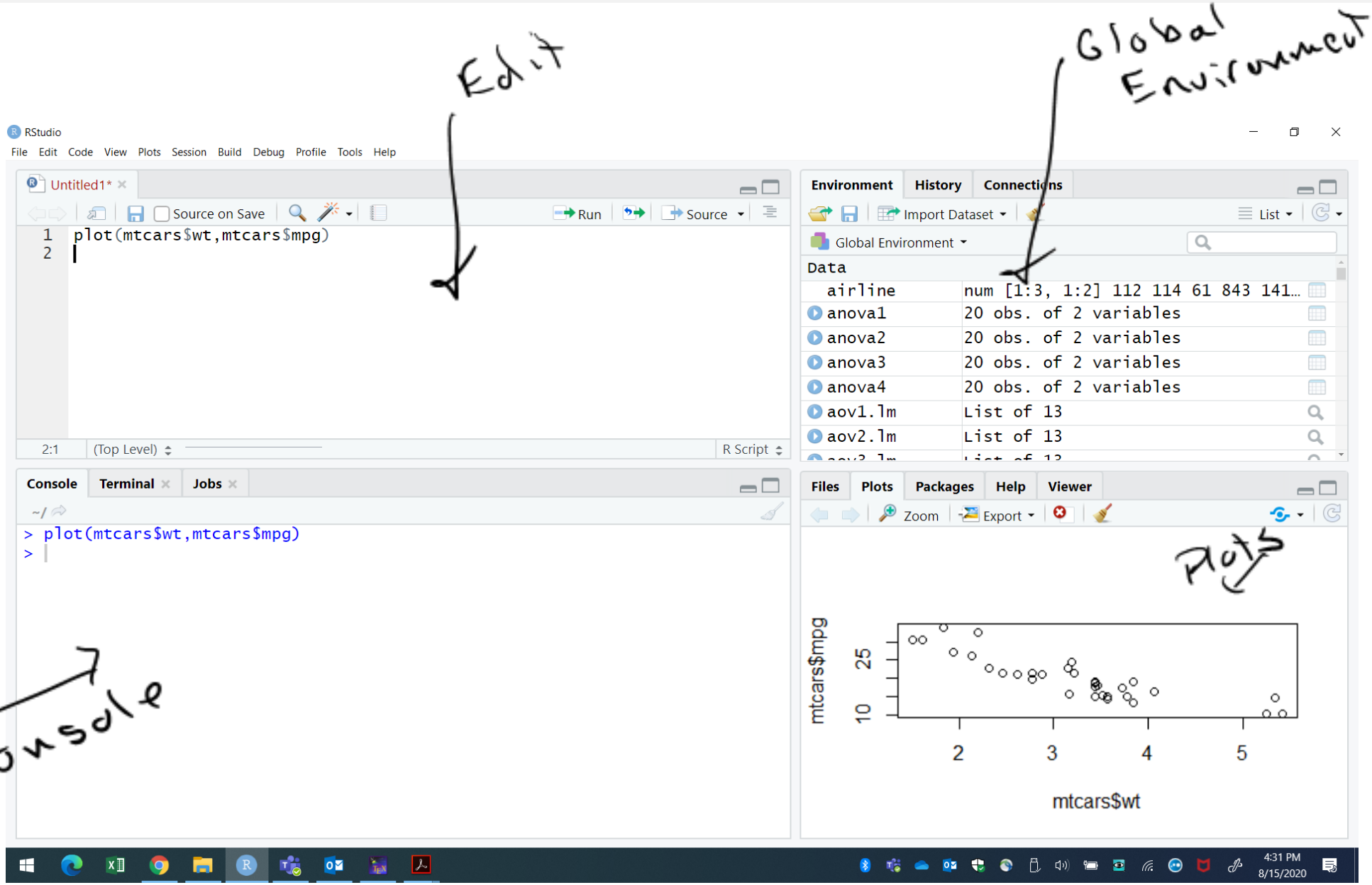
- R has become very popular over the past decade.
- It is an *open* source
- It is free
- Powerful enough to implement all of the methods discussed in this class
- Optional packages
- R is the language of choice for academic statisticians
- New approaches often become available in R years before they are implemented in commercial packages

Rstudio

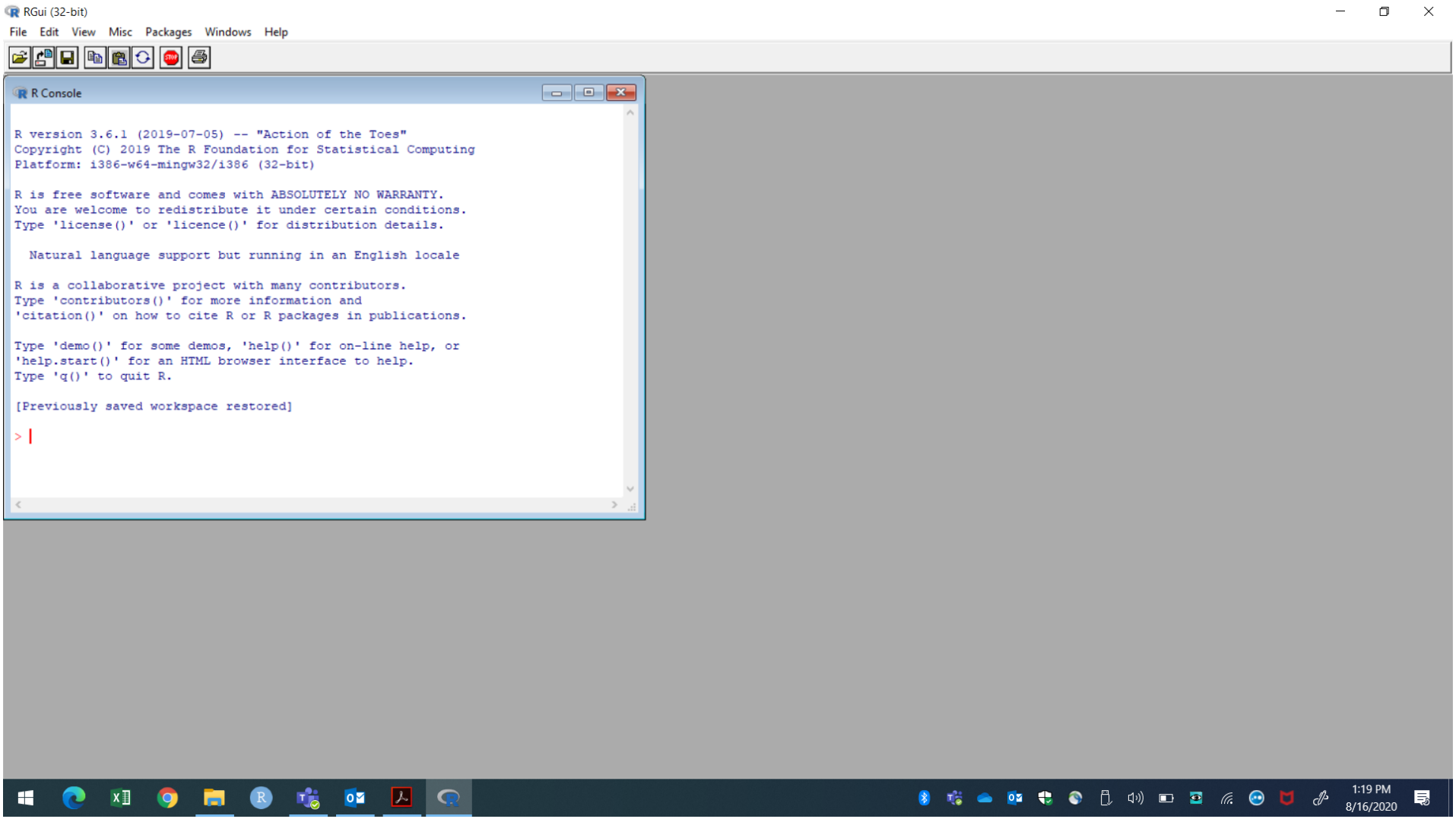
- In this class I will use Rstudio
- It is highly recommended that all users of R work in Rstudio
- Rstudio is an interface that provides both assistance for novices as well as productivity tools for experienced users.
- The Rstudio opens four windows:
  - One for editing code
  - A window for the console to execute R code
  - One track to the variables that are defined in the workspace
  - The fourth to display graphical images
- Rstudio is available through rstudio.cloud

Source: Applied Multivariate Statistics with R

Rstudio Windows



R Window


RGui (32-bit)
File Edit View Misc Packages Windows Help
R Console
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Previously saved workspace restored]
>
1:19 PM
8/16/2020

Example

- Suppose we want to predict miles per gallon (mpg) for automobiles based on certain values.
  - *cyl* Number of cylinders
  - *disp* Displacement (cu.in.)
  - *hp* Gross horsepower
  - *wt* Weight (1000 lbs)
  - *am* Transmission (0 = automatic, 1 = manual)
- This data set is in `R` called *mtcars*.
- At this time open up `Rstudio`

Lab Questions

Open `Rstudio` and select file → New File → R script answer the following questions

1. Type in the script ?mtcars and click Run. What year was this data was extracted from?
   a) 2019
   b) 2000
   c) 1984
   d) 1974

2. Type in the script dim(mtcars) then click Run the first number is the number of observations (rows) the second number is the number of variables (columns). How many observations?
   a) 11
   b) 32
   c) 352
   d) 43

n = 32 = # of observation
p = # of variables = 11

Lab Questions

Type in the script head(mtcars) then click Run.

3. How many rows appear?
   a) 32
   b) 16
   c) 6
   d) 2

4. How many cylinders are in the Hornet Sportabout?
   a) 4
   b) 6
   c) 8
   d) This car is not on the list.

Lab Questions

Lets compare the Weight of a car (*wt*) with the *mpg* by a plot.

5. In the script type in plot(wt,mpg), click Run. Describe this plot
   a) Positive, linear
   b) Negative, linear
   c) No relationship
   d) I got an error

6. In the script type in plot(mtcars$wt,mtcars$mpg), click Run. Describe this plot
   a) Positive, linear
   b) Negative, linear
   c) No relationship
   d) I got an error

To refer to a variable, we must type the data set and the variable name joined with a $ symbol. Alternatively, we can use the attach() function in order to tell R to make the variables in this data frame available by name. In the script window type in attach(mtcars) click Run.

Lab Questions

In the script window type in cyl = as.factor(mtcars$cyl) and click Run. Since the number of cylinders is numeric, R recognizes these values as continuous or quantitative. However, these should be categorical (factors).

7. Type in the script window plot(cyl,mpg) click Run, what plot do you see?
   a) Bar plot
   b) Boxplot
   c) Scatterplot
   d) Histogram
   e) I got an error

8. Type in the script window pairs(~mpg+disp+hp+wt,mtcars) click Run, what do you see in the graph window?
   a) Several scatterplots
   b) One scatterplot
   c) A histogram
   d) I got an error

Lab Questions

In the script window type summary(mtcars$mpg).

9. What is the mean of *mpg*?
   a) 15.43
   b) 19.20
   c) 20.09
   d) 22.80

If you want to save this script you can select File → save as . . . then save this were you want. It will save it with .R (R file).

What is Statistical Learning?

- **Statistical learning** refers to a vast set of tools for understanding data.
- These tools can be classified as *supervised* or *unsupervised*.
  - **Supervised statistical learning** involves building a statistical model for predicting, or estimating, an output based on one or more inputs.
  - **Unsupervised statistical learning** involves inputs but no supervising output.

Source: "An Introduction to Statistical Learning with Applications in R", page 1

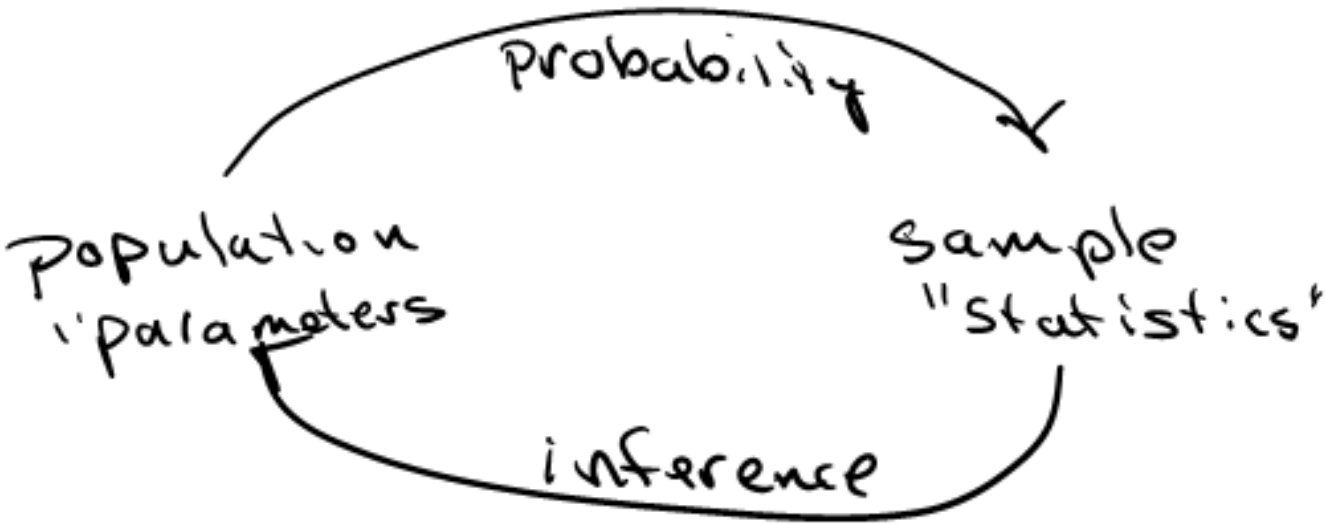
Goal of Statistical Learning

The main goal of statistical learning theory is to provide a framework for studying the problem of inference, that is of gaining knowledge, making predictions, making decisions or constructing models from a set of data. This is studied in a statistical framework, that is there are assumptions of statistical nature about the underlying phenomena (in the way the data is generated). [1]

[1] *Introduction to Statistical Learning Theory, O. Bousquet, S. Boucheron, and G. Lugosi*

What is Statistical Inference?

- A **statistical inference** aims at learning characteristics of the population from a sample
- The population characteristics are *parameters*
- The sample characteristics are *statistics*



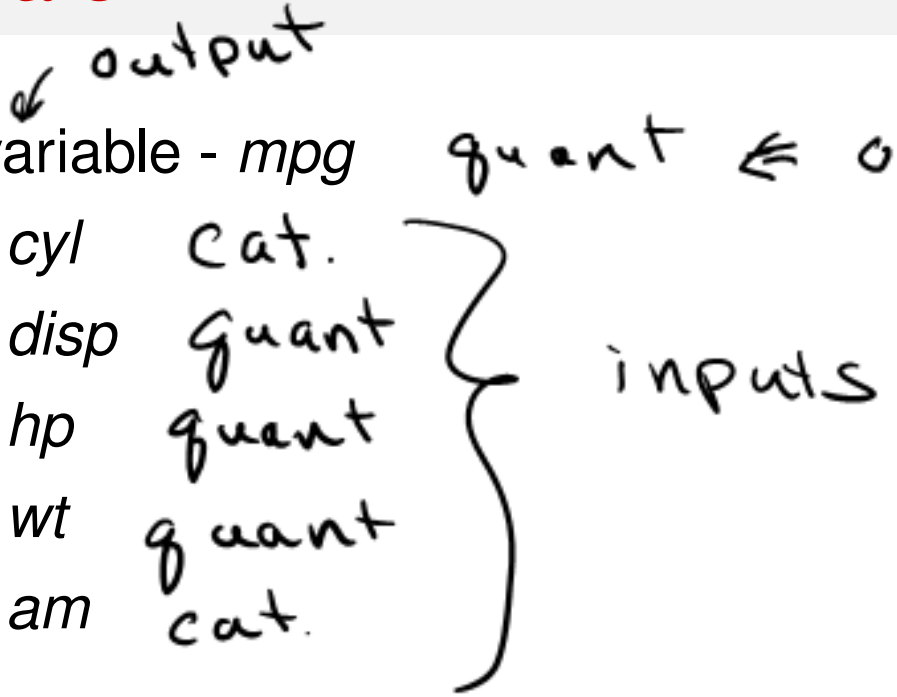
The Characteristics

In order to use the correct inference we need to know what type of characteristics we have from the data.

- Are these characteristics quantitative or categorical?
- Do we have a response variable (output or dependent variable) and factors (input, independent variable or predictor)?

Example *mtcars*

- Response variable - *mpg*
- Variable 1 - *cyl*
- Variable 2 - *disp*
- Variable 3 - *hp*
- Variable 4 - *wt*
- Variable 5 - *am*



Example 2

`https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236`

- variable 1 - DAY.OF.MONTH
- variable 2 - DAY.OF.WEEK
- variable 3 - CARRIER
- variable 4 - ORIGIN airport where the flight left
- variable 5 - DEST, airport where the flight landed
- variable 6 - DEP.DELAY, number of minutes that the departure was delayed, if negative left early
- variable 7 - ARR.DELAY, number of minutes that the landing was delayed, if negative came in early
- variable 8 - DISTANCE

1500 flights randomly sampled from the month of May 2018.

Two Analysis

- Given response and predictor - supervised learning
  - Regression - quantitative variables
  - Classification -categorical variable
- Given only factors without a response variable - unsupervised learning
  - Clustering

General Approach For Supervised Learning

- Let Y be the response (dependent variable).
- Let $X = (X_1, X_2, \ldots, X_p)$ be p different predictors (independent) variables.
- We assume there is some sort of relationship between X and Y, which can be written in the general form

$$Y = f(X) + \epsilon$$

- Statistical leaning refers to a set of approaches for estimating f.

Reasons for Estimating f

- Prediction: we want to predict Y, using $\hat{Y} = \hat{f}(X)$.
  - $\hat{f}$ is often treated as a **black box**.
  - The black box means that we are not typically concerned about the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.
- Inference: we want to know how Y is affected as X changes.
  - In this situation we wish to estimate f.
  - Thus $\hat{f}$ cannot be considered as a black box because we do want to know the exact form of $\hat{f}$.

Prediction $Y = f(X) + \epsilon$

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible error**
  - $\hat{f}$ is not a perfect estimate for f, and this inaccuracy will introduce some error.
  - We can potentially improve the accuracy of $\hat{f}$ by using the most appropriate statistical learning technique to estimate f.
- **Irreducible error**
  - However, even if it were possible to form a perfect estimate for f, so that our estimated response took the form $\hat{Y} = f(X)$, our prediction would still have some error in it!
  - This is because Y is also a function of ϵ, which, by definition, cannot be predicted using X.
  - Therefore, variability associated with ϵ also affects the accuracy of our predictions.
  - No matter how well we estimate f, we cannot reduce the error introduced by ϵ.

Inference Questions

If we are interested in inference we are asking we are interested in answering the the following questions:

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Types of Problems

- **Regression** problem is when we are the response is a *continuous* or *quantitative* output value.
- **Classification** problem is when the response is a *categorical* or *qualitative* output.

Regression or Classification?

In the following examples do would we use Regression methods or Classification methods? Also are we most interested in prediction or inference? What is the sample size (n) and the number of variables (p)?

1. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect the CEO salary.

2. We are considering launching a new product and wish to know whether it will be a success or failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price and 10 other variables.

How Do We Estimate f?

- The goal is to apply a statistical learning method to the training data in order to estimate the unknown function of f.
- Using a model-based approach, called **parametric**, with assumptions about the model.
  1. We make an assumption about the function form or shape of f.
  2. We need a procedure that uses training data to fit or train the model.
- No assumptions about the model is called a **non-parametric** method.
  - Non-parametric method seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.
  - **Advantage**: they have the potential to accurately fit a wider range of possible shapes for f.
  - **Disadvantage**: a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

Parametric Method

Parametric methods involve a two-step model-based approach.

1. We make an assumption about the functional form, or shape, of f. Then determine a model.
2. After a model has been selected, we need a procedure that uses the *training* data to fit or train the model.
   - The training data are observations used to train or teach our method how to estimate f.
   - Let x_{ij} represent the value of the jth predictor for observation i, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$.
   - Let y_i be the response variable for the ith observation.
   - Then the training data consist of $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

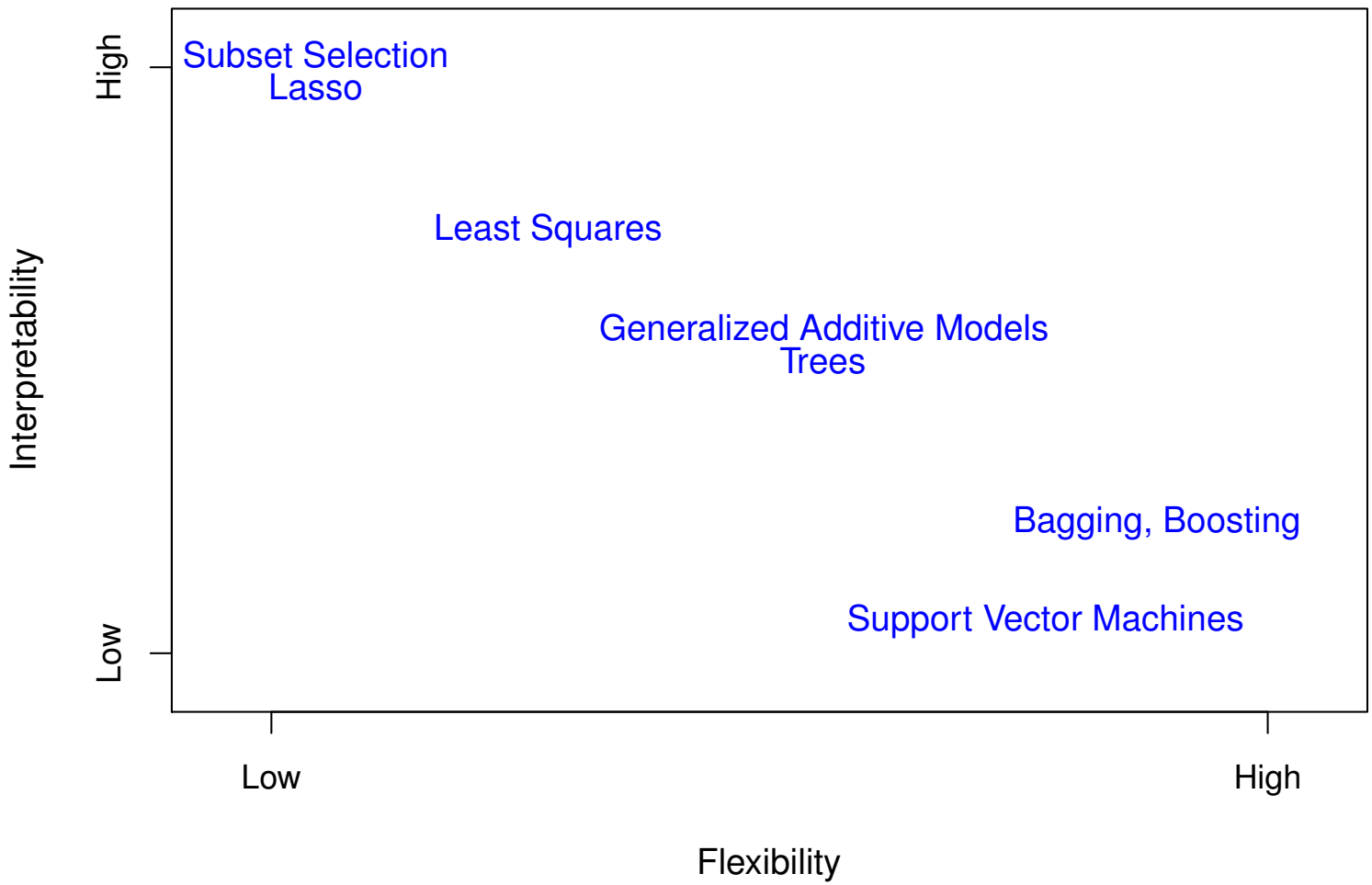
Flexibility vs. Interpretation

Why choose to use a more restrictive method instead of a very flexible approach?

- If we are mainly interested in inference, then restrictive models are much more interpretable. For example, when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the relationship between Y and $X_1, X_2, \ldots, X_p$.

- Very flexible approaches, such as the splines and the boosting methods can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.

Flexibility vs. Interpretation



Supervised and Unsupervised

Most statistical learning problems fall into one of two categories:

1. Supervised - for each observation of the predictor variables x, there is an associated response measurement y. Example: regression

2. Unsupervised - we observe variables x, but no associated response, y. Example: Cluster analysis

Lab Questions

10. What type of statistical learning problem is the *mtcars* example?
    a) Supervised regression
    b) Supervised classification
    c) Unsupervised

output = mpg