Lecture 1
Introduction to Statistical Learning
STAT 452, Spring 2021

What is Statistical Learning?

- Statistical learning refers to a vast set of tools for modeling and understanding complex data sets
- Two types of statistical learning problems:
  - **Supervised learning**: goal is to build a model that predicts an output variable using one or more input variables
  - **Unsupervised learning**: given input variables, but the output variable is not specified; one goal is to separate the data into relatively distinct groups (cluster analysis)
- The focus of this class will be on supervised learning.

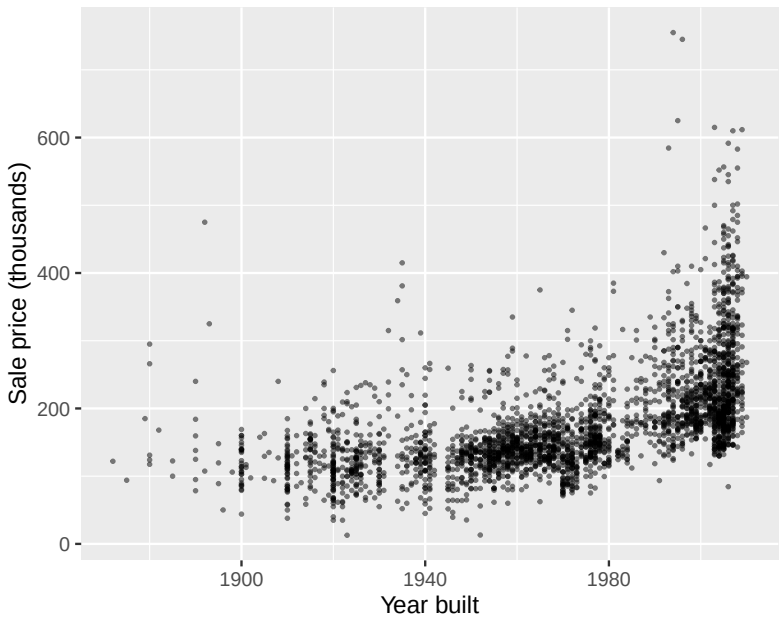
What is Statistical Learning?

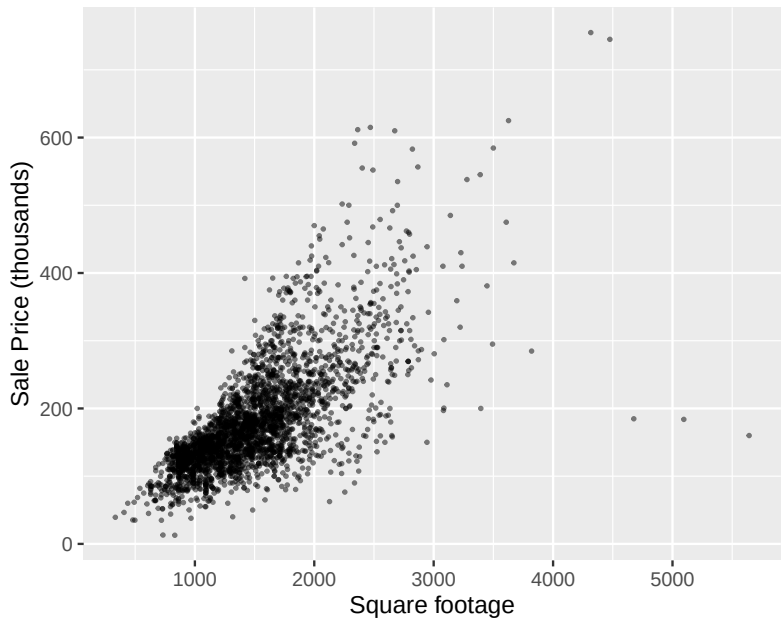
Some examples of statistical learning problems:

- Predict the price of stock in 6 months from now, on the basis of company performance and economic data
- Identify the numbers in a handwritten Zip code, from a digitized image
- Predict the condition of a stream based variables obtained from a Geographic Information System (GIS)
- Segment customers based on common attributes or purchasing behavior for targeted marketing
- Predict whether an email message is junk mail (spam)

Motivating Example

Data collected on the sales price for 2,930 residential properties, along with the year built and square footage.





Motivating Example

- Here sales price is the response variable (output) that we want to predict, which we refer to symbolically as Y.
- Year built and square footage are two predictor variables (inputs), which we refer to symbolically as X_1 and X_2.
- A major goal of statistical learning is to use the data to estimate a function $f(X_1, X_2)$ that predicts the Y as accurately as possible.

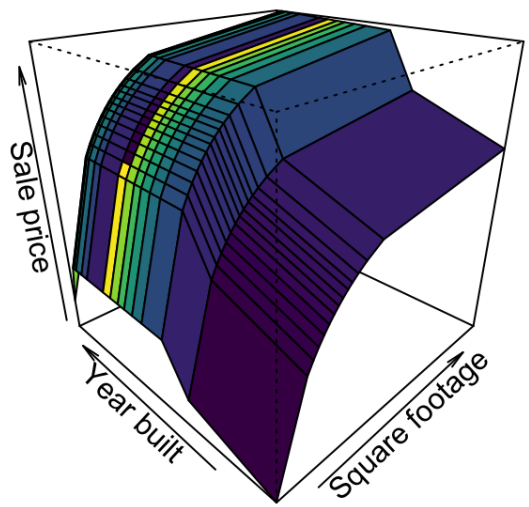


Figure: Plot that illustrates home sale price as a function of year built and square footage

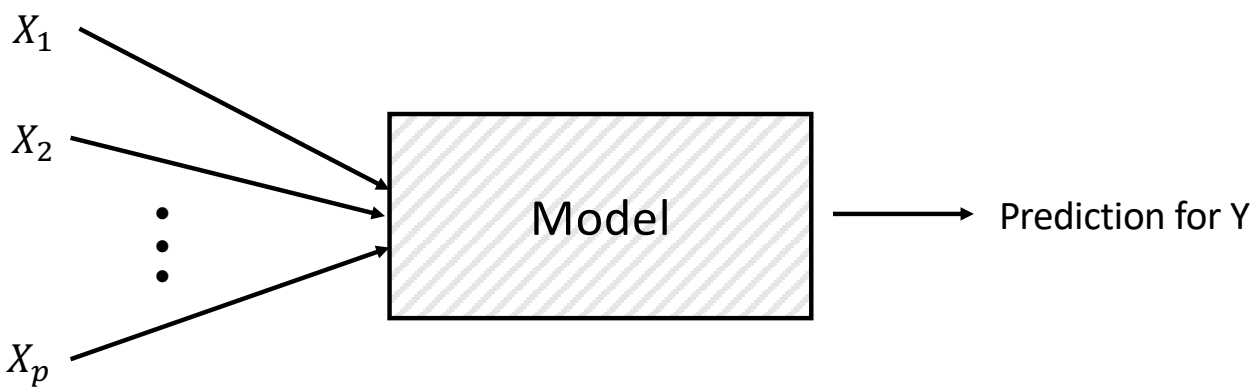
Obectives

Two main objectives for (supervised) statistical learning:

- **Prediction**: Make accurate predictions for future or new values of the response variable Y using predictors $X_1, X_2, \cdots, X_p$.
- **Inference**: Understand the relationship between the response variable Y and predictors $X_1, X_2, \cdots, X_p$. Some questions we may seek to answer:
  - Which predictors are associated with the response? Which predictors are most important?
  - Is the relationship between Y and the predictors linear or nonlinear? What mathematical function best describes the relationship?

Trade-off between Prediction Accuracy and Interpretability

- Often there are trade-offs between these two objectives. Simple models (e.g., linear regression) are easy to explain, while more complex models (e.g., random forests, neural nets) can potentially give better predictions.
- If all we care about is making accurate predictions, then we can think of our model as a so-called "black box". The inner-workings of the "black box" are not of interest, just that it makes good predictions.



Regression versus Classification Problems

- We tend to refer to problems with a numerical (quantitative) response as **regression** problems (e.g., predicting a person's age, height, or income).
- Problems involving a categorical (qualitative) response are often referred to as **classification** problems (e.g., predicting whether an email is `spam` or `not spam`; predicting `low`, `medium`, or `high` heath risk category for air quality).
- The use of this terminology can be a bit inconsistent. For instance, logistic regression is used for binary classification problems.

Label the following as either regression or classification problems:

- Predict a person's height based on their gender
- Predict a person's gender based on their height
- Predict whether or not a person has a particular disease based on their diet, weight, smoking status, and other health-related indicators
- Predict a person's salary based on their age, education, and occupation
- Identify handwritten digits (0-9) from a image based on pixel intensity

Topics

In this class, we will focus on the following statistical learning methods:

- Linear regression
- Logistic regression
- K-Nearest Neighbors
- Decision trees
- Random forests

Linear and logistic regression is also covered in STAT 432. However, the focus of this class will be on assessing prediction accuracy (cross-validation). Whereas the focus of STAT 432 is more on inference. A solid understanding of linear and logistic regression is also required before covering more sophisticated techniques.

Optional Topics

If we have time, we may also cover some of these additional topics:

- Generalized Additive Models
- Boosting
- Neural Nets
- Unsupervised Learning (Clustering Methods)

Software

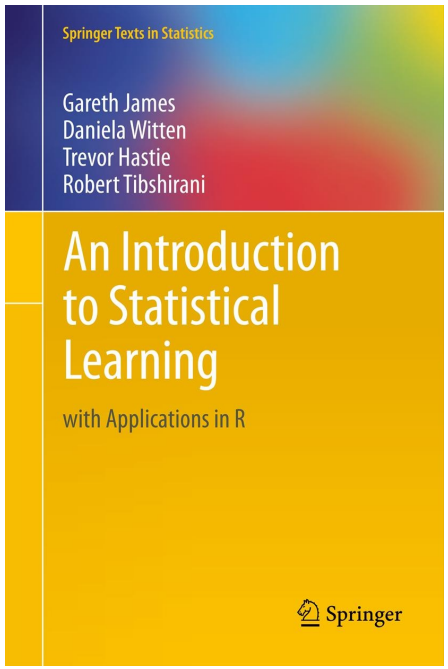
- We will be using R and R Studio in this class.
- R is a very good platform for doing machine learning. There are many packages that have been written for implementing the algorithms.
- It will be assumed that students have some prior computer programming experience, preferably with R.
- I can also provide students with access to R Studio Cloud, which will allow you to run R on an internet browser.

Textbooks

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

Free PDF version on Blackboard in the "Resources" folder.

This is one of the most well-known textbooks on statistical learning. It is written at the advanced undergraduate-level, and provides excellent descriptions of the different methods.



Textbooks

Bradley Boehmke and Brandon Greenwell. *Hands-On Machine Learning with R*. CRC Press, 2019.

Free online version: `https://bradleyboehmke.github.io/HOML/`

This is a more recent textbook on machine learning. It is well-written, and provides up-to-date R code for implementing machine learning methods, as well as some great example data sets.