0: Introduction to Statistical Learning

Data Sciences Institute
Linear Regression, Classification, and Resampling

Intro to course support team

- Technical Facilitator: Julia
- Learning Support Staff: Anjali
- Learning Support Staff: Vishakh
- Learning Support Staff: Dmytro

Welcome!

- So far, we've focused primarily on coding - now we explore the relationship between coding and statistics as this will allow us to answer questions such as "should we spend more of the advertising budget on TV or the internet"?
- This learning module will include definitions, mathematical concepts and approaches that may be new for most participants
- The learning curve will feel steep - this is expected - don't be hard on yourself if it takes time to sink in
- This module has recently been updated, with a greater emphasis on hands-on notebooks rather than mathematical theory.

Rules of Engagement

- You will be muted when you join the Zoom session. If you'd like to ask a question, please raise your hand to be unmuted!
- Questions are encouraged - ask as we go - this is your time to understand these concepts. However, please save *advanced* questions to office hours or work periods because sometimes they are excessively time consuming and may confuse most of your fellow participants who are beginners.
- If you have questions during live learning session, please try to ask them in the chats first where two of our Learning Supports will monitor the chats and answer questions there without interrupting the course. We will also pause and take time answering questions during the live course.

Let's navigate the LCR repo

https://github.com/UofT-DSI/LCR

Statistics and data science

What is Statistical Learning?

In simple terms, data science is all about using data to find useful information and insights.

Statistics is a big part of this process. It's the science of collecting and analyzing data to help us make smart decisions based on that data.

Statistics and data science

What is Statistical Learning?

Data science relies on two main things: statistics to understand the data and coding to handle it.

When we do data science, we use code to apply statistical methods and gain insights. We need to know both how to write the code and what the right statistical methods are to get meaningful results.

What is Statistical Learning?

Let’s imagine we work at a marketing agency. Our job is to help our client figure out how to spend their advertising budget to get the most sales.

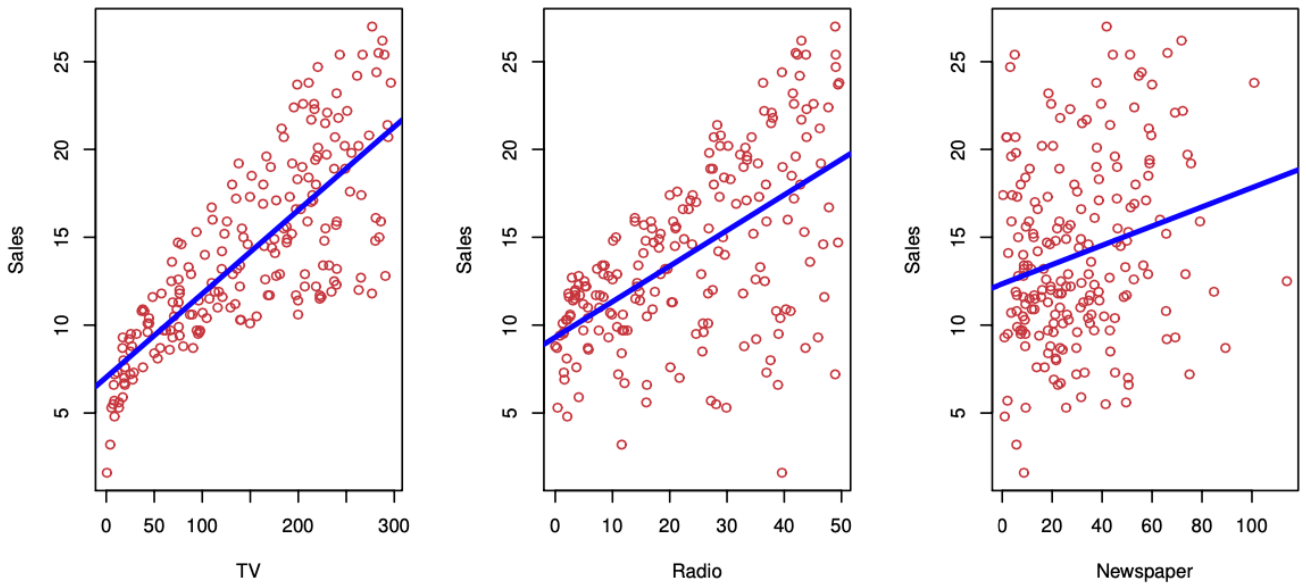
- The advertising budget for TV, radio, and newspapers are examples of things we measure, called predictors.
- The sales numbers are the response, or what we want to predict based on the budget.

What is Statistical Learning?

We want to understand how the budget for each advertising channel affects the sales. This is where statistical learning comes in.

We use a simple formula to show how the budget (predictor) might be related to sales (response), but this is not perfect because the world is complicated.

What is Statistical Learning?



The relationship between the advertising budgets and sales can be shown by a line of best fit, but there will always be some random error in our predictions. This random error shows that no model is perfect.

Statistical learning helps us estimate this line, even though it's not going to be perfect.

Types of Statistical Learning

Prediction vs Inference

There are two main reasons why we want to understand this relationship:

Prediction: We want to predict future sales based on the advertising budget.

Inference: We want to understand how changes in the advertising budget will affect sales.

Types of Statistical Learning

Prediction

In prediction, we focus on using what we know (the budget) to guess what will happen (sales).

Our goal is to make our guesses as accurate as possible, but we know that some errors will always be out of our control. We try to reduce those errors that we can control.

Types of Statistical Learning

Inference

In inference, we focus on understanding the relationship between the budget and sales.

We want to figure out which factors are the most important and how they affect sales. We also want to understand if the relationship is simple or complex.

Applying Statistical Learning

How do we estimate our outcome?

To estimate the relationship between the budget and sales, we typically split our data into two parts:

- **Training data:** Used to teach the model how to estimate the relationship.
- **Testing data:** Used to see how accurate the model is when predicting new data.

Applying Statistical Learning

Supervised vs Unsupervised Learning

- **Supervised learning** This is when we predict something (like sales) based on other information (like the advertising budget).
    - Examples of supervised learning models are linear regression and classification. These models are the primary focus of this learning module.
- **Unsupervised learning** This is when we try to understand the relationships between different pieces of information without predicting anything.
    - There is no response variable to predict, instead, the goal is to understand the relationship between variables or observations.
    - An example of this is clustering.

Applying Statistical Learning

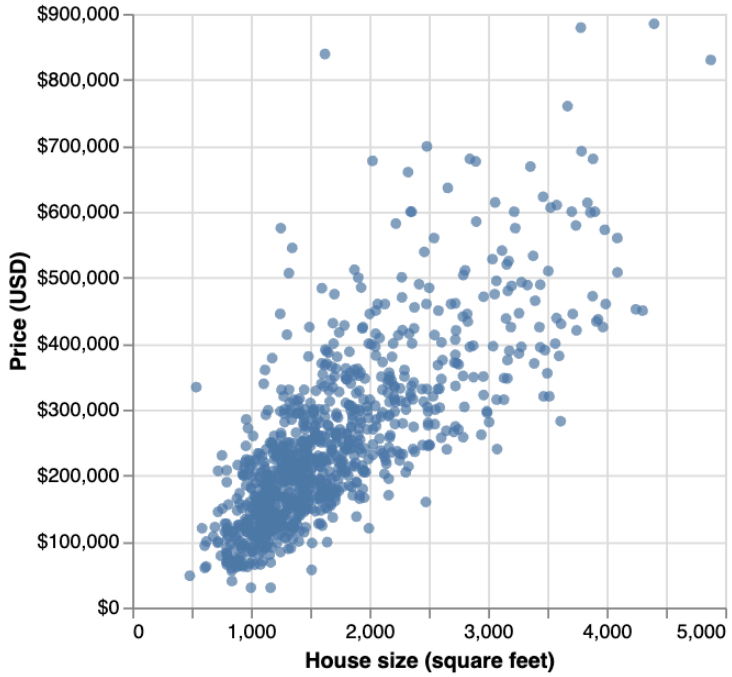
Regression vs Classification Problems

- Variables can be numbers (like age or salary) or categories (like yes/no or disease/healthy).
- ***Regression problems:*** We predict numerical values, like sales based on the budget.
- ***Classification problems:*** We predict categories, like whether someone will buy a product (yes or no).
- This is a bit of a simplification, but for our purposes, we can think of regression as predicting numerical values and classification as predicting categories.

Types of Plots

In upcoming modules, you'll dive deeper into plotting techniques, but for now, let's quickly review some common plots that will appear in this module:

- **Scatter plots**: display the relationship between two continuous variables
    - For example, relationship between house size and house price



- **Histograms**: display the distribution of a numeric variable's values as a series of bars, with each bar representing the count of values within a specific range.
    - For example, frequency of price per night ($)