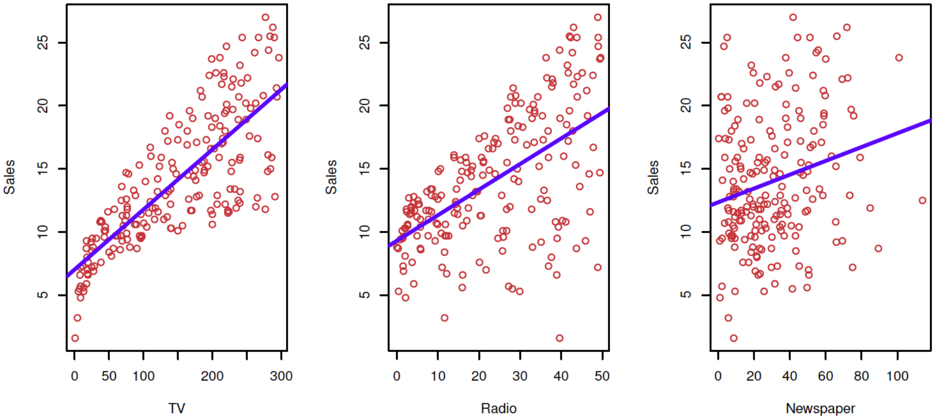
2.1 What is Statistical Learning?

Prof. Lauren Perry

Motivating Example

Goal: think about how to increase sales by investigating association between product advertising and sales.



Sales (thousands of units) for various advertising types in 200 different markets.

Some Terminology

- The client has no control over `sales`, so this is the *output variable*.
  - Also called the *response* or *dependent variable*.
- By contrast, `advertising` is the *input variable*.
  - Also called *predictors*, *independent variables*, *features*, or just *variables*.

Notation

Response: Y

Predictors: $X_1, X_2, \ldots, X_p$

Some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$:

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

where

- f is some fixed, unknown function
- ϵ is a random *error term* such that $\epsilon \perp\!\!\!\perp X$ and $\mathrm{E}(\epsilon) = 0$.

Motivating Example

With our `sales` (Y) and advertising data `TV` (X_1), `radio` (X_2), and `newspaper` (X_3), the model

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

- allows us to make predictions about `sales`.
- gives us insight into which components of $X = (X_1, X_2, X_3)$ are important to explaining Y.
- may give us information about how each component X_j affects Y (depending on complexity of f).

Estimating f: Prediction

Often we have a set of values X, but not Y.

We predict values of Y using

$$\hat{Y} = \hat{f}(X)$$

- We may or may not be concerned with the actual form of $\hat{f}$.

Prediction Accuracy

This depends on two quantities:

Reducible error: $\hat{f}$ is not a perfect estimate of f, but we can improve it by improving our estimating techniques.

Irreducible error: the variability associated with Y (which comes from the random variable ϵ), which we cannot reduce.

- We will always have this error, which is generally due to unmeasured (or unmeasurable) variables.

Prediction Accuracy

We can show that

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = [f(X) - \hat{f}(X)]^2 + Var(\epsilon)$$

reducible error $+$ irreducible error.

This course will focus on techniques to estimate f and minimize reducible error.

Inference

Sometimes, we are also interested in the association between Y and X.

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

This requires us to know the exact form of $\hat{f}$.

Inference

Examples:

- Which health measures are associated with heart disease risk?
- Which health measure is associated with the largest decrease in heart disease risk?
- How much of an increase in heart disease risk is associated with a given increase in number of cigarettes smoked per day?

Often, we want to do both inference *and* prediction for a given problem.

Estimating f: Parametric Methods

Step 1:

- Make an assumption about the functional form of f.
  - Example: f is linear in X:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Estimating f: Parametric Methods

Step 2:

- Use the training data to fit/train the model.
  - In the linear model

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

we must estimate the p coefficients β_1, β_2, …, β_p.

Estimating f: Parametric Methods

Parametric models reduce the problem of estimating f down to estimating a set of parameters.

- Pro: (significantly) simplifies the problem of estimating f.
- Con: the model chosen is probably not the true form of f.
  - We can improve our model by making it more flexible...
  - ...but more flexible models run the risk of *overfitting* (fitting the noise too closely).

We will learn about specific parametric models in the next couple chapters.

Estimating f: Nonparametric Methods

Nonparametric methods make no explicit assumptions about the functional form of f.

- Goal: estimate f so that it gets close to the data points without being too "wiggly" (without overfitting).
- Pro: can fit a much wider range of possible shapes for f.
- Con: often require a large number of observations to obtain accurate estimates.

We will learn more about nonparamteric methods in later chapters.

Accuracy versus Interpretability

If we have enough data, why would we ever use a less flexible parametric approach?

- Restrictive models tend to be much more interpretable (wrt inference).
  - More complex methods may make it very difficult to understand the relationship between a predictor and the response.

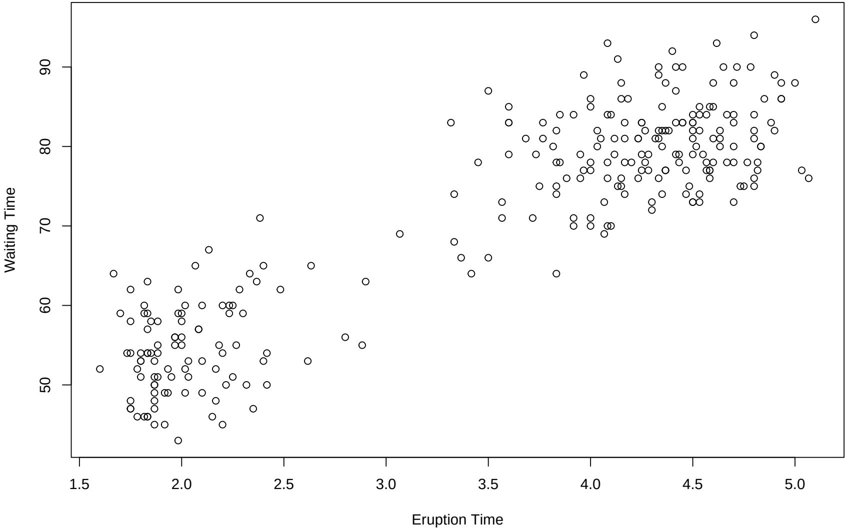
If we're only interested in prediction, we may not have a strong reason to use a restrictive model.

- However, very flexible models can have a tendency to overfit the data, so we shouldn't always go straight to the most flexible approach possible!

Supervised versus Unsupervised Learning

- In *supervised learning* problems, we use a set of predictor variables X to predict some response variable Y using $Y = f(X) + \epsilon$.
- In *unsupervised learning* problems, we have a set of variable X, but we don't have a response variable.
  - Want to understand the relationships between the variables/observations.
  - Often we use some kind of *clustering* algorithm to group the observations.

A Simple Example of Unsupervised Learning



Regression versus Classification

Broadly,

- *Regression* models refer to models with a *quantitative* (numeric) response Y.
- *Classification* models refer to models with a *qualitative* (categorical) response Y.

There is some nuance here, but we will worry about that on a case-by-case basis.