Applied Statistics - Notes

260236

March 2024

Preface

Every theory section in these notes has been taken from two sources:

- An Introduction to Statistical Learning [1]
- Applied Multivariate Statistical Analysis (sixth edition). [2]

About:

GitHub repository

Contents

1 Sample Geometry

1.1 The Geometry of the Sample

A single **multivariate observation** is the **collection of measurements on p different variables taken on the same item or trial**. If n **observations** have been obtained, the entire data set can be placed in an $n \times p$ array (or matrix), also called **data frame**:

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

Each **row** of $\mathbf{X}$ represents a **multivariate observation**. Since the entire data frame is often one particular realization of what might have been observed, we say that the data frame are a **sample of size n from a p-variate "population"**. The sample then consists of n measurements, each of which has p components.

Look at the matrix, n measurements (rows), each of which has p components (columns). In mathematics, each n row contains p columns and vice versa.

The data frame can be plotted in two different ways:

1. p-dimensional scatter plot, where the rows represent n points in p-dimensional space;

2. Geometrical representation, p vectors in n-dimensional space.

1.1.1 Scatter plot

For the **p-dimensional scatter plot**, the rows of $\mathbf{X}$ represent n points in p-dimensional space:

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \begin{array}{l} \leftarrow \text{1st (multivariate) observation} \\ \\ \\ \leftarrow n\text{th (multivariate) observation} \end{array} \tag{2}$$

The row vector $\mathbf{x}_j'$, representing the jth observation, contains the coordinates of a point. The **scatter plot** of n points in p-dimensional space **provides information** on the **locations and variability of the points**.

Note: when p (dimensional space) is greater than 3, the **scatter plot** representation cannot actually be graphed. Yet the consideration of the data as n points in p dimensions provides **insights that are not readily available from algebraic expressions**.

1.1.2 Geometrical representation

The alternative **geometrical representation** is constructed by considering the data as p **vectors in n-dimensional space**. Here we take the elements of the columns of the data frame to be the coordinates of the vectors:

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [\mathbf{y}_1 \mid \mathbf{y}_2 \mid \cdots \mid \mathbf{y}_p] \tag{3}$$

Then the **coordinates** of the first point $\mathbf{y}_1 = [x_{11}, x_{21}, \ldots, x_{n1}]$ **are the** n **measurements** on the first variable.

In general, the ith point $\mathbf{y}_i = [x_{11}, x_{21}, \ldots, x_{n1}]$ is determined by the n-tuple of all measurements on the ith variable.

Geometrical representations usually **facilitate understanding** and lead to further insights. The ability to **relate algebraic expressions to the geometric concepts** of length, angle and volume is therefore **very important**.

1.1.3 Geometrical interpretation of the process of finding a sample mean

Before starting the explanation, you need to understand a few things.

- The **length** of a vector $\mathbf{x}' = [x_1, x_2, \ldots, x_n]$ with n components is defined by:

  $$L_x = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{4}$$

  Multiplication of a vector $\mathbf{x}$ by a scalar c changes the length:

  $$\begin{aligned} L_{cx} &= \sqrt{c^2 \cdot x_1^2 + c^2 \cdot x_2^2 + \cdots + c^2 \cdot x_n^2} \\ &= |c| \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \\ &= |c|\, L_x \end{aligned}$$

  So, for example, in $n = 2$ dimensions, the vector:

  $$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

  The length of $\mathbf{x}$, written L_x, is defined to be:

  $$L_x = \sqrt{x_1^2 + x_2^2}$$

- Another important concept is **angle**. Consider two vectors in a plane and the angle θ between them: The value θ can be represented as the

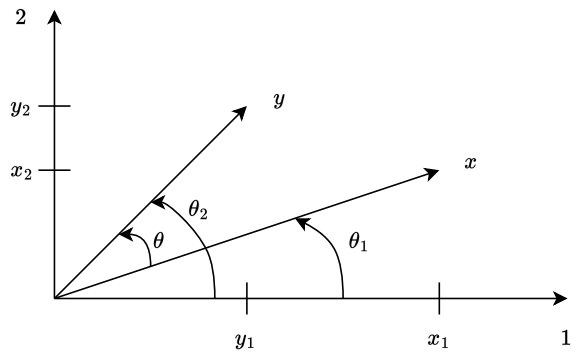


Figure 1: The angle θ between $\mathbf{x}' = [x_1, x_2]$ and $\mathbf{y}' = [y_1, y_2]$.

difference between the angles θ_1 and θ_2 formed by the two vectors and the first coordinate axis. Since, by definition:

$$\cos(\theta_1) = \frac{x_1}{L_x} \quad \cos(\theta_2) = \frac{y_1}{L_y}$$
$$\sin(\theta_1) = \frac{x_2}{L_x} \quad \sin(\theta_2) = \frac{y_2}{L_y}$$
$$\cos(\theta) = \cos(\theta_2 - \theta_1) = \cos(\theta_2)\cos(\theta_1) + \sin(\theta_2)\sin(\theta_1)$$

The angle θ between the two vectors $\mathbf{x}' = [x_1, x_2]$ and $\mathbf{y}' = [y_1, y_2]$ is specified by:

$$\cos(\theta) = \cos(\theta_2 - \theta_1) = \left(\frac{y_1}{L_y}\right)\left(\frac{x_1}{L_x}\right) + \left(\frac{y_2}{L_y}\right)\left(\frac{x^2}{L_x}\right) = \frac{x_1 y_1 + x_2 y_2}{L_x L_y} \tag{5}$$

- With the angle equation 5, it's convenient to introduce the **inner product** of two vectors:

$$\mathbf{x}\mathbf{y}' = x_1 y_1 + x_2 y_2$$

So let us rewrite:

  - The **length** equation 4:

$$\mathbf{x}'\mathbf{x} = x_1 x_1 + x_1 x_1 = x_1^2 + x_2^2 \longrightarrow L_x = \sqrt{x_1^2 + x_2^2} \Longrightarrow L_x = \sqrt{\mathbf{x}'\mathbf{x}} \tag{6}$$

  - The **angle** equation 5:

$$\cos(\theta) = \frac{x_1 y_1 + x_2 y_2}{L_x L_y} \Longrightarrow \cos(\theta) = \frac{\mathbf{x}'\mathbf{y}}{L_x L_y}$$

    And using the rewritten length equation:

$$\cos(\theta) = \frac{\mathbf{x}'\mathbf{y}}{L_x L_y} \Longrightarrow \cos(\theta) = \frac{\mathbf{x}'\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x}} \cdot \sqrt{\mathbf{y}'\mathbf{y}}}$$

- The **projection** (or shadow) of a vector $\mathbf{x}$ on a vector $\mathbf{y}$ is:

$$\frac{(\mathbf{x}'\mathbf{y})}{\mathbf{y}'\mathbf{y}}\mathbf{y} = \frac{(\mathbf{x}'\mathbf{y})}{L_y}\frac{1}{L_y}\mathbf{y} \tag{7}$$

Where the vector $\frac{1}{L_y}\mathbf{y}$ has unit length. The **length of the projection** is:

$$\frac{|\mathbf{x}'\mathbf{y}|}{L_y} = L_x \left|\frac{\mathbf{x}'\mathbf{y}}{L_x L_y}\right| = L_x \left|\cos(\theta)\right| \tag{8}$$

Where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$:

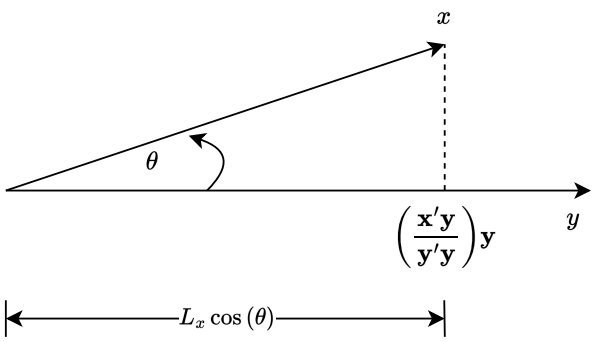


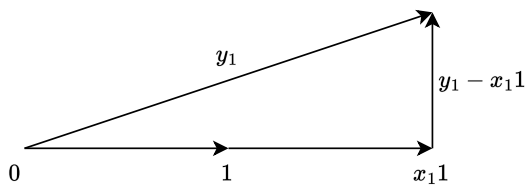
Figure 2: The projection of $\mathbf{x}$ on $\mathbf{y}$.

Start by defining the $n \times 1$ vector $\mathbf{1}'_n = [1, 1, \ldots, 1]$. The vector $\mathbf{1}$ forms equal angles with each of the n coordinates axes, so the vector $\left(\frac{1}{\sqrt{n}}\right)\mathbf{1}$ has unit length in the equal-angle direction. Consider the vector $\mathbf{y}'_i = [x_{1i}, x_{2i}, \ldots, x_{ni}]$. The projection of $\mathbf{y}_i$ on the unit vector $\left(\frac{1}{\sqrt{n}}\right)\mathbf{1}$ is:

$$\mathbf{y}'_i\left(\frac{1}{\sqrt{n}}\mathbf{1}\right)\frac{1}{\sqrt{n}}\mathbf{1} = \frac{x_{1i} + x_{2i} + \cdots + x_{ni}}{n}\mathbf{1} = \overline{x}_i\mathbf{1} \tag{9}$$

Although it may seem like a complex equation at first glance, it is nothing more than the mean! In fact, the **sample mean** $\overline{\mathbf{x}}_i = \frac{(x_{1i} + x_{2i} + \cdots + x_{ni})}{n} = \frac{\mathbf{y}'_i\mathbf{1}}{n}$ corresponds to the multiple of $\mathbf{1}$ required to give the projection of $\mathbf{y}_i$ onto the line determined by $\mathbf{1}$.

Furthermore, using the projection, you can obtain the **deviation** (**mean corrected**). For each $\mathbf{y}_i$ we have the decomposition:



Where $\overline{x}_i\mathbf{1}$ is perpendicular to $y_i - \overline{x}_i\mathbf{1}$. The **deviation**, or **mean corrected**, vector is:

$$\mathbf{d}_i = \mathbf{y}_i - \overline{x}_i\mathbf{1} = \begin{bmatrix} x_{1i} - \overline{x}_i \\ x_{2i} - \overline{x}_i \\ \vdots \\ x_{ni} - \overline{x}_i \end{bmatrix} \tag{10}$$

The **elements** of $\mathbf{d}_i$ are the **deviations of the measurements on the ith variable from their sample mean**.

Using the length rewritten with inner product (equation 6) and the deviation (equation 10), we obtain:

$$L^2_{\mathbf{d}_i} = \mathbf{d}_i'\mathbf{d}_i = \sum_{j=1}^{n} (x_{ji} - \overline{x}_i)^2 \tag{11}$$

$$(\text{Length of deviation vector})^2 = \text{sum of squared deviations}$$

From the sample standard deviation, we see that the **squared length is proportional to the variance** of the measurements on the ith variable. Equivalently, the **length is proportional to the standard deviation**. So longer vectors represent more variability than shorter vectors.

Furthermore, for any two deviation vectors $\mathbf{d}_i$ and $\mathbf{d}_k$:

$$\mathbf{d}_i'\mathbf{d}_k = \sum_{j=1}^{n} (x_{ji} - \overline{x}_i)(x_{jk} - \overline{x}_k) \tag{12}$$

And with a few mathematical operations, we can get it:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \cos(\theta_{ik}) \tag{13}$$

Where the **cosine** of the angle is the **sample correlation coefficient**. Note: s_{ik} is the **sample covariance**:

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n} (x_{ji} - \overline{x}_i)(x_{jk} - \overline{x}_k) \qquad i = 1, 2, \ldots, p, \quad k = 1, 2, \ldots, p \tag{14}$$

Thus:

- If the two deviation vectors have **nearly the same orientation**, the sample correlation will be close to 1;

- If the two vectors are **nearly perpendicular**, the sample correlation will be approximately zero;
- If the two vectors are oriented in **nearly opposite directions**, the sample correlation will be close to -1.

1.2 Generalized Variance

Before starting the explanation, you need to understand what is a sample variance.

A **sample variance** is defined as:

$$s_k^2 = s_{kk} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{jk} - \overline{x}_k)^2 \qquad k = 1, 2, \ldots, p \tag{15}$$

With a single variable, the **sample variance is often used to describe the amount of variation in the measurements on that variable**. When p variables are observed on each unit, the variation is described by the **sample variance-covariance matrix**:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} = \left\{ s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ji} - \overline{x}_i)(x_{jk} - \overline{x}_k) \right\} \tag{16}$$

The sample covariance matrix contains p variances and $\frac{1}{2}p(p-1)$ potentially different covariances. Sometimes it's desirable to **assign a single numerical value for the variation expressed by S**. One choice for a value is the determinant of $\mathbf{S}$, which reduces to the usual sample variance of a single characteristic when $p = 1$. This determinant is called the **generalized sample variance**:

$$\text{Generalized sample variance} = \det(\mathbf{S}) = |\mathbf{S}| \tag{17}$$

2 Statistical Learning

2.1 Introduction

Suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the general form:

$$Y = f(X) + \varepsilon \tag{18}$$

Where ε is an **error term**, which is **independent** of X and has **mean zero**. The function f represents the **systematic information** that X provides about Y. The **function** f that connects the input variables to the output variable **is in general unknown**.

Example 1

For **example**, on the left-hand panel of figure 3, a plot `income` versus `years of education` for 30 individuals in the Income data set.

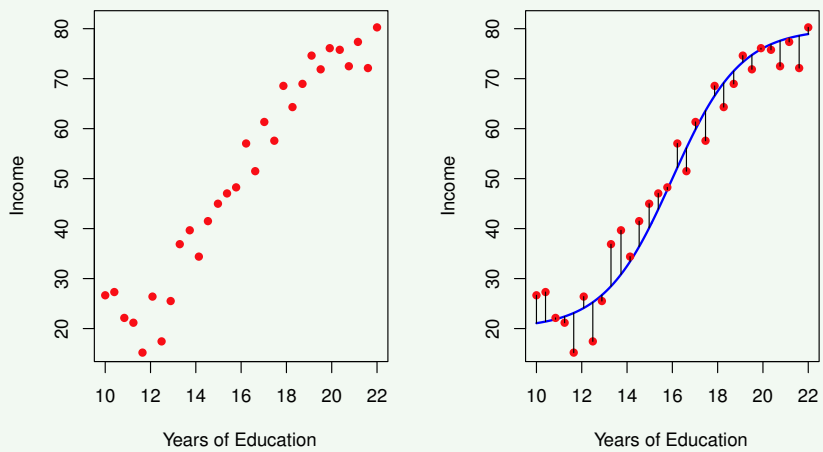


Figure 3: The `Income` data set. [1]

As you can see, the plot suggests that one might be able to predict `income` using `years of education`. Since `Income` is a simulated data set, the function f is known and is shown by the blue curve in the right-hand panel. The **vertical lines** represent the **error terms** ε. We note that some of the 30 observations lie above the blue curve and some lie below it; overall, the **errors have approximately mean zero**.

In essence, **statistical learning refers to a set of approaches for estimating** f. In this chapter we outline some of the key theoretical concepts that arise in estimating f.

2.2 Why Estimate f (systematic information provided by a predictor about a quantitative response)?

There are two main reasons that we may wish to estimate f: **prediction** and **inference**.

2.2.1 Prediction

In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term ε averages to zero, we can predict Y using:

$$\hat{Y} = \hat{f}(X) \tag{19}$$

- $\hat{f}$ represents our **estimate for f**
- $\hat{Y}$ represents **prediction** for Y

The function $\hat{f}$ is often treated as a **black box**, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that **it yields accurate predictions for Y**.

Example 2

As an **example**, suppose that:

- $X_1, \ldots, X_p$ are **characteristics of a patient's blood sample** that can be easily measured in a lab.
- Y is a variable encoding the **patient's risk for a severe adverse reaction to a particular drug**.

It is natural to seek to predict Y using X, since we can then avoid giving the drug in question to patients who are at high risk of an adverse reaction. That is, patients for whom the estimate of Y is high.

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities: **reducible error** and **irreducible error**.

- In general, $\hat{f}$ will not be a perfect estimate for f, and this **inaccuracy** will introduce some error. This is a **reducible error** because we can potentially **improve the accuracy of $\hat{f}$ by using the most appropriate statistical learning technique to estimate f**.
- 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 (error term), which, by definition, cannot be predicted using X. Therefore, variability associated with ε also affects the accuracy of our predictions. This is the **irreducible error**, because **no matter how well we estimate f, we cannot reduce the error introduced by ε**.

The real question is: *why is the irreducible error larger than zero?* Well, the quantity ε may contain unmeasured variables that are useful in predicting Y: since we don't measure them, f cannot use them for its prediction. The quantity ε may also contain unmeasurable variation.

Example 3

For **example**, the risk of an adverse reaction might vary for a given patient on a given day, depending on manufacturing variation in the drug itself or the patient's general feeling of well-being on that day.

Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed, so that the only variability comes from ε (error term). Then, it's easy to show that:

$$\begin{aligned} E\left(Y - \hat{Y}\right)^2 &= E\left[f(X) + \varepsilon - \hat{f}(X)\right]^2 \\ &= \underbrace{\left[f(X) - \hat{f}(X)\right]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible}} \end{aligned} \tag{20}$$

- $\left[f(X) - \hat{f}(X)\right]^2$ represents the **squared difference between the predicted and actual value of Y**
- $E\left(Y - \hat{Y}\right)^2$ represents the **average**, or **expected value**
- $\text{Var}(\varepsilon)$ represents the **variance associated with the error term ε**

The focus of this course is on *techniques* for estimating f with the aim of **minimizing the reducible error**. It is important to keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for Y. Unfortunately, this bound is almost always unknown in practice.

Example 4

Consider a company that is interested in conducting a direct-marketing campaign.
The *goal* is to identify individuals who are likely to respond positively to a mailing, based on observations of demographic variables measured on each individual.
In this case:

- The demographic variables serve as *predictors*;
- Response to the marketing campaign (either positive or negative) serves as the *outcome*.

The company is <u>not</u> interested in obtaining a deep understanding of the relationships between each individual predictor and the response; instead, the company simply **wants to accurately predict the response using the predictors**.
This is an example of **modeling for prediction**.

2.2.2 Inference

We are often interested in understanding the association between Y (quantitative response) and $X_1, \ldots, X_p$ (p-predictors). In this situation we wish to estimate f (systematic information), but our goal is not necessarily to make predictions for Y. Now it's obviously that $\hat{f}$ cannot be treated as a black box, because we need to know its exact form. In this setting, one may be interested in **answering the following questions**:

- ***Which predictors are associated with the response?*** It is often the case that only a small fraction of the available predictors are substantially associated with Y. So, **identifying** the few **important predictors among a large set of possible variables can be extremely useful**.
- ***What is the relationship between the response and each predictor?*** Larger values of the predictor are associated with larger values of Y. Other predictors may have the opposite relationship. The relationship between the response and the given predictor may **depend** on:
  - The **complexity** of f;
  - The **values of the other predictors**.
- ***Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?*** Historically, **most methods** for estimating f **have** taken **linear form**. But often the true relationship is more complicated, in which case a **linear model may not provide an accurate representation** of the relationship between the input and the output variables.

Example 5

Modeling the brand of a product that a customer might purchased based on variables such as:

- Price
- Store
- Location
- Discount levels
- Competition price

And so forth. In this situation one might really be most interested in the **association between each variable and the probability of purchase**. For instance, *to what extent is the product's price associated with sales?*
This is an example of **modeling for inference**.

Example 6

Consider the following figure:

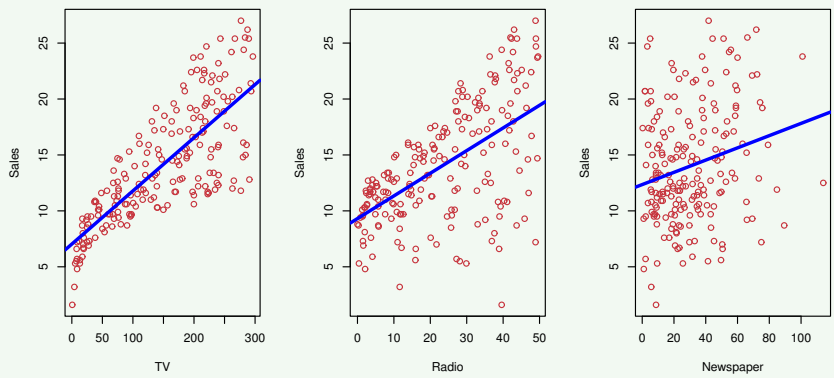


Figure 4: The `Advertising` data set. The plot displays `sales`, in thousands of units, as a function of `TV`, `radio`, and `newspaper` budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of `sales` to that variable. In other words, each blue line represents a simple model that can be used to predict `sales` using `TV`, `radio`, and `newspaper`, respectively.

One may be interested in answering questions such as:

- *Which media are associated with sales?*
- *Which media generate the biggest boost in sales?*
- *How large of an increase in sales is associated with a given increase in TV advertising?*

This situation falls into the **inference model**.

2.2.3 Difference between prediction and inference

Example 7

In a real estate setting, one may seek to relate values of homes to inputs such as:

- Crime rate
- Zoning
- Distance from a river
- Air quality
- Schools
- Income level of community
- Size of houses

And so forth. In this case one might be interested in the association between each individual input variable and housing price. For instance, *how much extra will a house be worth if it has a view of the river?* This is an **inference problem**.

But attention! Alternatively, one may simply be interested in predicting the value of a home given its characteristics: *is this house under or over valued?* And this is a **prediction problem**.

So, as you can see from the example, the difference between a prediction problem and an inference problem is so small. A problem can change its nature because the ultimate goal is also changing.

2.3 How do we estimate f?

We will always assume that we have observed a set of n different data points. For example, in figure 3 at page 10 we observed $n = 30$ data points. These observations are called **training data** because we will **use these observations to train, or teach, our method how to estimate f**.

Let:

- x_{ij} represent the value of the jth predictor, or input, for observation i, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$
- y_i represent the response variable for the ith observation.

Then, our training data consist of:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

Where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

Our goal is to **apply a statistical learning method to the training data in order to estimate the unknown function f**. In other words, we want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observations (X, Y). Most statistical learning methods for this task can be characterized as either **parametric** or **non-parametric**.

2.3.1 Parametric Methods

The **parametric methods** involve a two-step model-based approach:

1. Select a model.

   (a) **Make an assumption about the functional form**, or shape, of f. For **example**, one very simple assumption is that f is linear in X:

   $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{21}$$

   This is a **linear model** (that will be discussed in the future). Once we have assumed that f is linear, **the problem of estimating f is greatly simplified**. Instead of having to estimate an entirely arbitrary p-dimensional function $f(X)$, one only needs to **estimate the $p+1$ coefficients** $\beta_0, \beta_1, \ldots, \beta_p$.

2. Use training data to fit/train the model.

   (b) After a model has been selected, we need a **procedure that uses the training data to fit the model** or **train the model**. In the case of the linear method, we need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$. So, we want to find values of these parameters such that:

   $$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

   The most **common approach to fitting** the (linear) model is referred to as (**ordinary**) **least squares** (that will be discussed in the future). However, the least squares is one of many possible ways to fit the linear model.

The parametric model-based reduces the problem of estimating f down to one of **estimating a set of parameters**. In fact, assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters in the linear model, than it is to fit an entirely arbitrary function f.

⚠ Potential disadvantage

The **model** we choose will **usually not match the true unknown form of f**. If the chosen model is **too far** from the true f, then our **estimate will be poor**.

✔ Possible (partial) solution

We can try to address this problem by **choosing flexible models** that can **fit many different possible functional forms for f**. But fitting a more flexible model **requires estimating a greater number of parameters**.

These more complex models (**flexible models**) can lead to a phenomenon known as **overfitting** the data, which essentially means **they follow the errors**, or **noise**, **too closely** (these issues are discussed throughout this course).

2.3.2 Non-Parametric Methods

The **non-parametric** methods do not make explicit assumptions about the functional form of f. Instead they seek an **estimate of f that gets as close to the data points as possible without being too rough or wiggly**.

✔ Major advantage over parametric approaches

By avoiding the assumption of a particular functional form for f, non-parametric approaches have the **potential to accurately fit a wider range of possible shapes** for f. Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well.

⚠ Disadvantage

Since non-parametric approaches do not reduce the problem of estimating f to a small number of parameters, **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.

References

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2013.

[2] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.

Index