Elements of Nonparametric Statistics

Nicholas Henderson

2020-01-08

Contents

Preface

This book will serve as the main source of course notes for Biostatistics 685/Statistics 560, Winter 2020.

Chapter 1

Introduction

1.1 What is Nonparametric Statistics?

What is Parametric Statistics?

- Parametric models refer to probability distributions that can be fully described by a fixed number of parameters that do not change with the sample size.
- Typical examples include
  - Gaussian
  - Poisson
  - Exponential
  - Beta
- Could also refer to a regression setting where the mean function is described by a fixed number of parameters.

What is Nonparametric Statistics?

- It is difficult to give a concise, all-encompassing definition, but nonparametric statistics generally refers to statistical methods where there is not a clear parametric component.
- A more practical definition is that nonparametric statistics refers to flexible statistical procedures where very few assumptions are made regarding the distribution of the data or the form of a regression model.

- The uses of nonparametric methods in several common statistical contexts are described in Sections 1.3 - 1.7.

1.2 Outline of Course

This course is roughly divided into the following 5 categories.

1. **Nonparametric Testing**

- Rank-based Tests
- Permutation Tests

1. **Estimation of Basic Nonparametric Quantities**

- The Empirical Distribution Function
- Density Estimation

1. **Nonparametric Confidence Intervals**

- Bootstrap
- Jacknife

1. **Nonparametric Regression Part I (Smoothing Methods)**

- Kernel Methods
- Splines
- Local Regression

1. **Nonparametric Regression Part II (Machine Learning Methods)**

- Decision Trees/CART
- Ensemble Methods

1.3 Example 1: Nonparametric vs. Parametric Two-Sample Testing

Suppose we have data from two groups. For example, outcomes from two different treatments.

- **Group 1 outcomes**: $X_1, \ldots, X_n$ an i.i.d (independent and identically distributed) sample from distribution function F_X. This means that

$$F_X(t) = P(X_i \leq t) \quad \text{for any } 1 \leq i \leq n$$

- **Group 2 outcomes**: $Y_1, \ldots, Y_m$ an i.i.d. sample from distribution function F_Y.

$$F_Y(t) = P(Y_i \leq t) \quad \text{for any } 1 \leq i \leq n$$

- To test the impact of a new treatment, we usually want to test whether or not F_X differs from F_Y in some way. This can be stated in hypothesis testing language as

$$\begin{aligned} H_0 &: F_X = F_Y \quad (\text{ populations are the same}) \\ H_A &: F_X \neq F_Y \quad (\text{ populations are different}) \end{aligned} \tag{1.1}$$

Parametric Tests

- Perhaps the most common parametric test for (1.1) is the **t-test**. The t-test assumes that

$$F_X = \text{Normal}(\mu_x, \sigma^2) \quad \text{and} \quad F_Y = \text{Normal}(\mu_y, \sigma^2) \tag{1.2}$$

- Under this parametric assumption, the hypothesis test (1.1) reduces to

$$H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_A : \mu_x \neq \mu_y \tag{1.3}$$

- The standard t-statistic (with a pooled estimate of σ^2) is the following

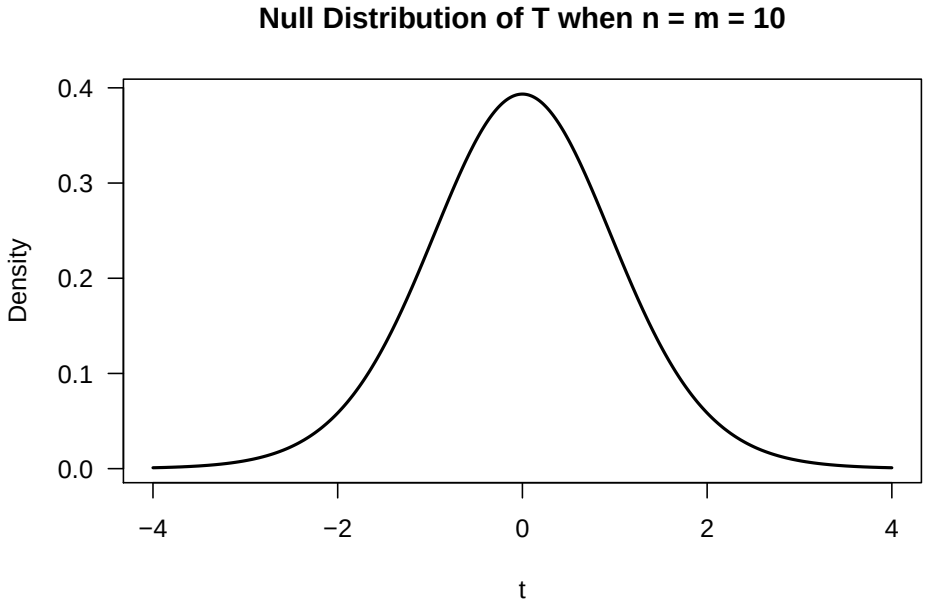
$$T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}, \tag{1.4}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$ are the group-specific sample means and s_p^2 is the pooled estimate of σ^2

$$s_p^2 = \frac{1}{m+n-2}\Big\{ \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2 \Big\} \tag{1.5}$$

- The t-test is based on the **null distribution** of T - the distribution of T under the null hypothesis.

- Under the assumption of normality, the null distribution of T is a t distribution with $n + m - 2$ degrees of freedom.



- Notice that the null distribution of T depends on the parametric assumption that both $F_X = \text{Normal}(\mu_x, \sigma^2)$ and $F_Y = \text{Normal}(\mu_y, \sigma^2)$. Appealing to the Central Limit Theorem, one could argue that is a quite reasonable assumption.
- In addition to using the assumption that $F_X = \text{Normal}(\mu_x, \sigma^2)$ and $F_Y = \text{Normal}(\mu_y, \sigma^2)$, we used this parametric assumption (at least implicitly) in the formulation of the hypothesis test itself because we assumed that any difference between F_X and F_Y would be fully described by difference in μ_x and μ_y.
- So, in a sense, you are using the assumption of normality twice in the construction of the two-sample t-test.

Nonparametric Tests

- Two-sample nonparametric tests are meant to be "distribution-free". This means the null distribution of the test statistic does not depend on any parametric assumptions about the two populations F_X and F_Y.
- Many such tests are based on **ranks**. The distribution of the ranks under the assumption that $F_X = F_Y$ do not depend on the form of F_X (assuming F_X is continuous).
- Also, the statements of hypotheses tests for nonparametric tests should not rely on any parametric assumptions about F_X and F_Y.
- For example, $H_A : F_X \neq F_Y$ or $H_A : F_X \geq F_Y$.

- Nonparametric tests usually tradeoff power for greater robustness.
- In general, if the parametric assumptions are correct, a nonparametric test will have less power than its parametric counterpart.
- If the parametric assumptions are not correct, parametric tests might have inappropriate type-I error control or lose power.

1.4 Example 2: Nonparametric Estimation

- Suppose we have n observations $(X_1, \ldots, X_n)$ which are assumed to be i.i.d. (independent and identically distributed). The distribution function of X_i is F_X.
- Suppose we are interested in estimating the entire distribution function F_X rather than specific features of the distribution of X_i such as the mean or standard deviation.
- In a **parametric** approach to estimating F_X, we would assume the distribution of X_i belongs to some parametric family of distributions. For example,
  - $X_i \sim \text{Normal}(\mu, \sigma^2)$
  - $X_i \sim \text{Exponential}(\lambda)$
  - $X_i \sim \text{Beta}(\alpha, \beta)$

- If we assume that $X_i \sim \text{Normal}(\mu, \sigma^2)$, we only need to estimate 2 parameters to fully describe the distribution of X_i, and the number of parameters will not depend on the sample size.
- In a nonparametric approach to characterizing the distribution of X_i, we need to instead estimate the entire distribution function F_X or density function f_X.
- The distribution function F_X is usually estimated by the **empirical distribution function**

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t), \tag{1.6}$$

  where $I()$ denotes the indicator function. That is, $I(X_i \leq t) = 1$ if $X_i \leq t$, and $I(X_i \leq t) = 0$ if $X_i > t$.
- The empirical distribution function is a discrete distribution function, and it can be thought of as an estimate having n "parameters.

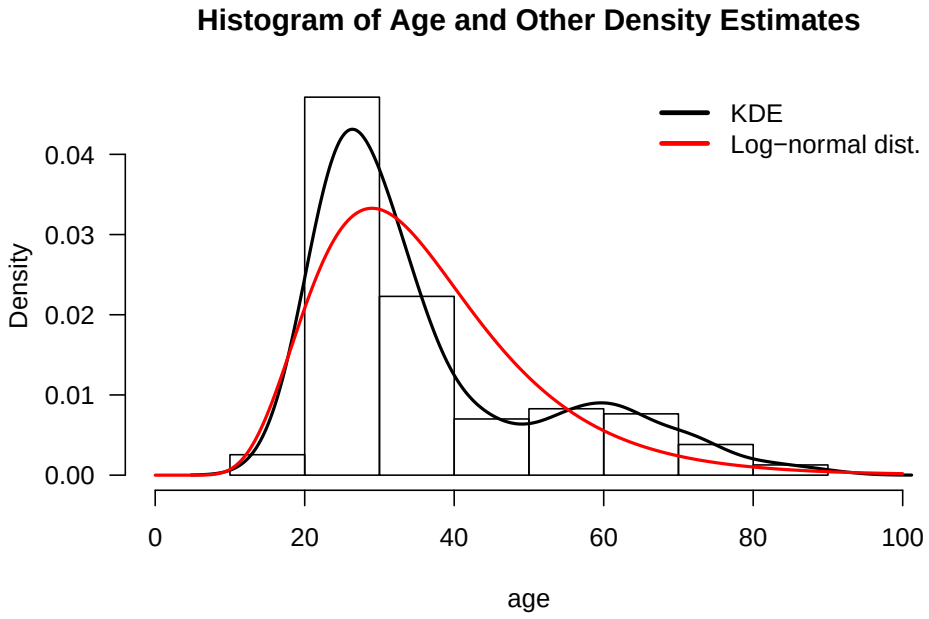
- The density function of X_i is often estimated by a kernel density estimator (KDE). This is defined as

$$\hat{f}_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\Big(\frac{t - X_i}{h_n}\Big). \tag{1.7}$$

- $K()$ - the kernel function

- h_n - the bandwidth

- The KDE is a type of smoothing procedure.



1.5 Example 3: Confidence Intervals

- Inference for a wide range of statistical procedures is based on the following argument

$$\hat{\theta}_n \text{ has an approximate Normal}\left(\theta, \widehat{\text{Var}}(\hat{\theta}_n)\right) \text{ distribution} \tag{1.8}$$

- Above, $\hat{\theta}_n$ is an estimate of a parameter θ, and $\widehat{\text{Var}}(\hat{\theta}_n)$ is an estimate of the variance of $\hat{\theta}_n$.
- $se_n = \sqrt{\widehat{\text{Var}}(\hat{\theta}_n)}$ is usually referred to as the **standard error**.
- 95% confidence intervals are reported using the following formula

$$[\hat{\theta}_n - 1.96se_n, \hat{\theta}_n + 1.96se_n] \tag{1.9}$$

- Common examples of this include:
  1. $\hat{\theta}_n = \bar{X}_n$.

  In this case, appeals to the Central Limit Theorem would justify approximation (1.8). The variance of $\hat{\theta}_n$ would be σ^2, and the standard error would typically be $se_n = \hat{\sigma}/\sqrt{n}$.

  2. $\hat{\theta}_n$ = Maximum Likelihood Estimate of θ.

In this case, asymptotics would justify the approximate distribution $\hat{\theta}_n \sim \text{Normal}(\theta, \frac{1}{nI(\theta)})$, where $I(\theta)$ denotes the Fisher information. The standard error in this context is often $se_n = \{nI(\hat{\theta}_n)\}^{-1/2}$.

- Confidence intervals using (1.8) rely on a parametric approximation to the sampling distribution of the statistic $\hat{\theta}_n$.
- Moreover, even if one wanted to use something like (1.8), working out standard error formulas can be a great challenge in more complicated situations.

- The **bootstrap** is a simulation-based approach for computing standard errors and confidence intervals.
- The bootstrap does not rely on any particular parametric assumptions and can be applied in almost any context (though bootstrap confidence intervals can fail to work as desired in some situations).
- Through resampling from the original dataset, the bootstrap uses many possible alternative datasets to assess the variability in $\hat{\theta}_n$.

	OriginalDat	Dat1	Dat2	Dat3	Dat4
Obs. 1	0.20	0.20	0.80	0.20	0.30
Obs. 2	0.50	0.20	0.80	0.20	0.70
Obs. 3	0.30	0.30	0.50	0.80	0.20
Obs. 4	0.80	0.30	0.70	0.50	0.50
Obs. 5	0.70	0.70	0.20	0.30	0.20
theta.hat	0.50	0.34	0.60	0.40	0.38

- In the above example, we have 4 **boostrap replications** for the statistic $\hat{\theta}$:

$$\hat{\theta}^{(1)} = 0.34 \tag{1.10}$$
$$\hat{\theta}^{(2)} = 0.60 \tag{1.11}$$
$$\hat{\theta}^{(3)} = 0.40 \tag{1.12}$$
$$\hat{\theta}^{(4)} = 0.38 \tag{1.13}$$

- In the above example, the bootstrap standard error for $\hat{\theta}_n$ would be the

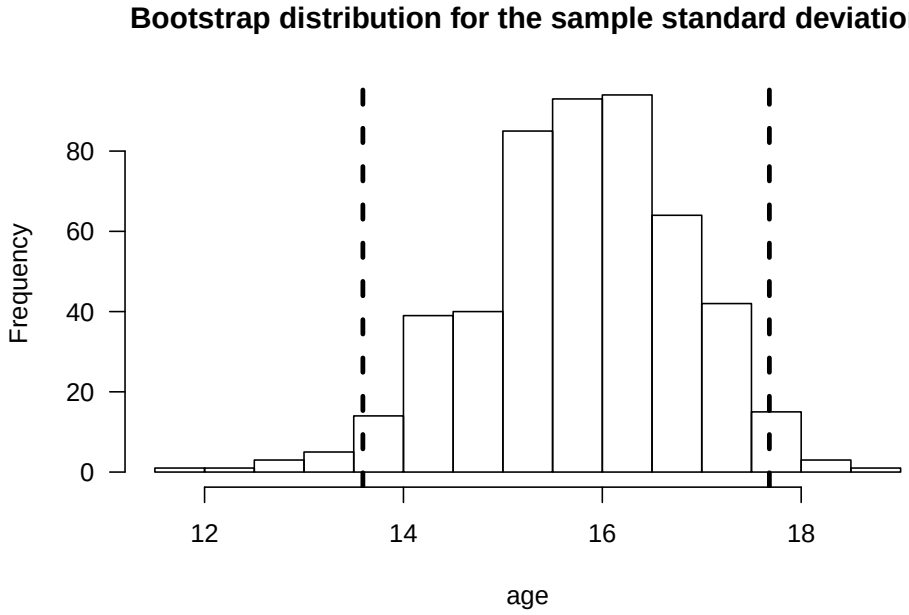


Figure 1.1: Bootstrap distribution of the sample standard deviation for the age variable from the kidney fitness data. Dasjed vertical lines are placed at the 2.5 and 97.5 percentiles of the bootstrap distribution.

standard deviation of the bootstrap replications

$$\begin{aligned} se_{boot} &= \Big(\frac{1}{3}\sum_{b=1}^{4}\{\hat{\theta}^{(b)} - \hat{\theta}^{(-)}\}^2\Big)^{1/2} \\ &= \Big((0.34-0.43)^2/3 + (0.60-0.43)^2/3 + (0.40-0.43)^2/3 + (0.38-0.43)^2/3\Big)^{1/2} \\ &= 0.116 \end{aligned} \tag{1.14}$$

where $\hat{\theta}^{(-)} = 0.43$ is the average of the bootstrap replications.

- One would then report the confidence interval $[\hat{\theta} - 1.96 \times 0.116, \hat{\theta} + 1.96 \times 0.116]$. In practice, the number of bootstrap replications is typically much larger than 4.

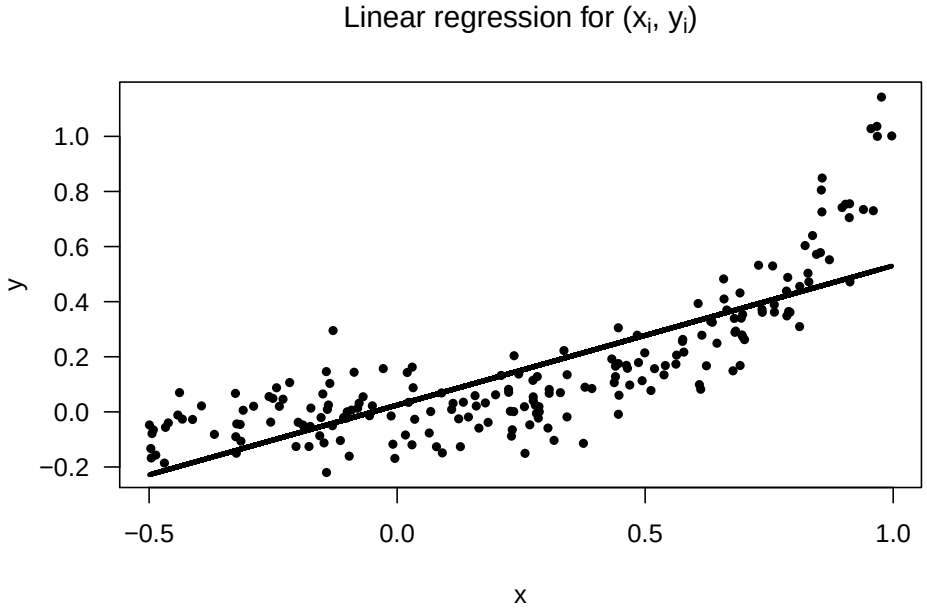
- It is often better to construct confidence intervals using the percentiles from the bootstrap distribution of $\hat{\theta}$ rather than use a confidence interval of the form: $\hat{\theta} \pm 1.96 \times se_{boot}$.

1.6 Example 4: Nonparametric Regression with a Single Covariate

- Regression is a common way of modeling the relationship between two different variables.
- Suppose we have n pairs of observations $(y_1, x_1), \ldots, (y_n, x_n)$ where y_i and x_i are suspected to have some association.
- Linear regression would assume that these y_i and x_i are related by the following

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1.15}$$

  with the assumption $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$ often made.
- In this model, there are only 3 parameters: $(\beta_0, \beta_1, \sigma^2)$, and the number of parameters stays fixed for all n.



- The nonparametric counterpart to linear regression is usually formulated in the following way

$$y_i = m(x_i) + \varepsilon_i \tag{1.16}$$

- Typically, one makes very few assumptions about the form of the mean function m, and it is not assumed m can be described by a finite number of parameters.
- There are a large number of nonparametric methods for estimating m.

- One popular method is the use of **smoothing splines**.

- With smoothing splines, one considers mean functions of the form

$$m(x) = \sum_{j=1}^{n} \beta_j g_j(x) \tag{1.17}$$

where $g_1, \ldots, g_n(x)$ are a collection of spline basis functions.

- Because of the large number of parameters in (1.17), one should estimate the basis function weights β_j through penalized regression

$$\text{minimize} \quad \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{n} \beta_j g_j(x_i) \Big)^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \Omega_{ij} \beta_i \beta_j \tag{1.18}$$

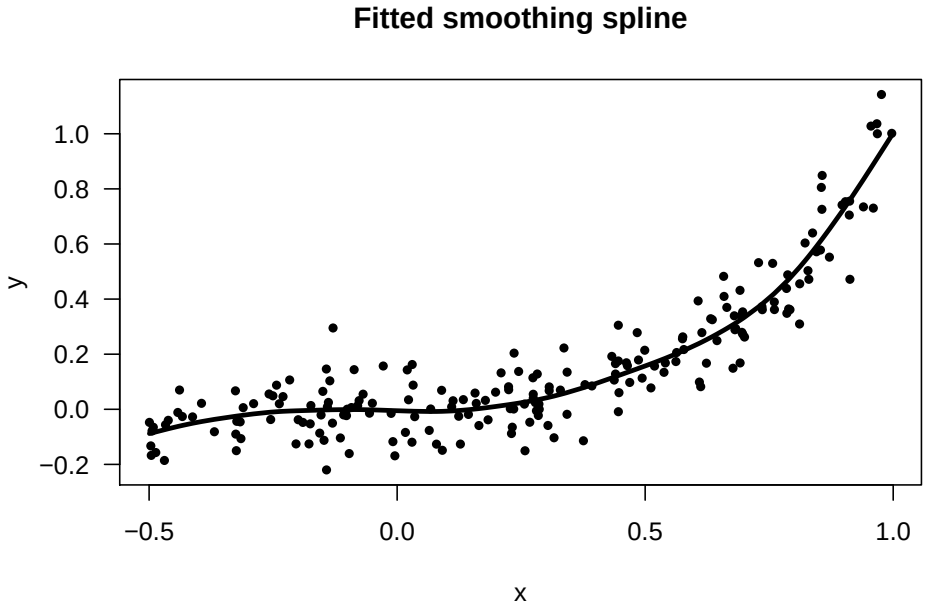
where $\Omega_{ij} = \int g_i''(t) g_j''(t) dt$.

- Using coefficient estimates $\hat{\beta}_1, \ldots, \hat{\beta}_n$ found from solving (1.17), the nonparametric estimate of the mean function is defined as

$$\hat{m}(x) = \sum_{j=1}^{n} \hat{\beta}_j g_j(x) \tag{1.19}$$

- While the estimation in (1.18) resembles parametric estimation for linear regression, notice that the number of parameters to be estimated will change with the sample size.

- Allowing the number of basis functions to grow with n is important. For a sufficiently large number of basis functions, one should be able to approximate the true mean function $m(x)$ arbitrarily closely.



1.7 Example 5: Classification and Regression Trees (CART)

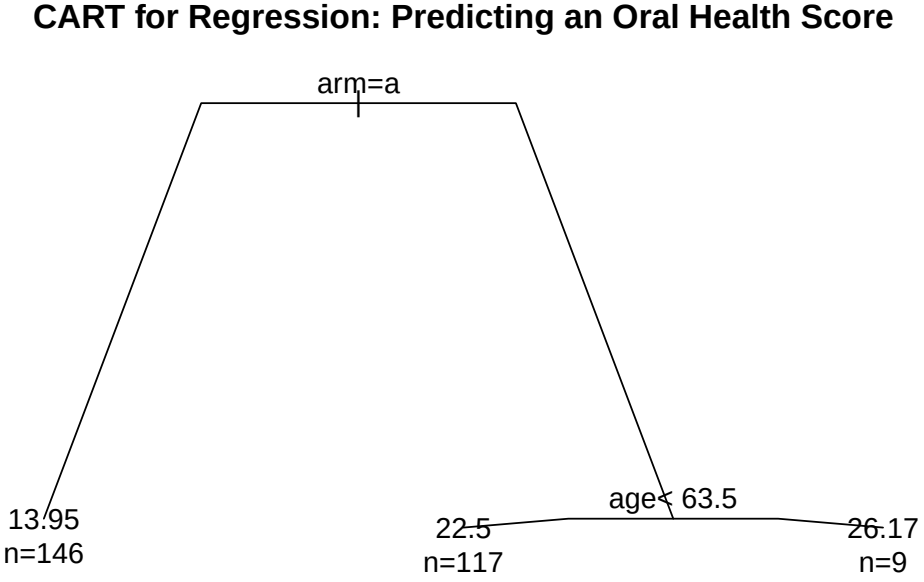
- Suppose we now have observations $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ where y_i is a continuous response and $\mathbf{x}_i$ is a p-dimensional vector of covariates.

- Regression trees are a nonparametric approach for predicting y_i from $\mathbf{x}_i$.

- Here, the regression function is a **decision tree** rather than some fitted curve.

- With a decision tree, a final prediction from a covariate vector $\mathbf{x}_i$ is obtained by answering a sequence of "yes or no" questions.

- When the responses y_i are binary, such trees are referred to as classification trees. Hence, the name: classification and regression trees (CART).



- Classification and regression trees are constructed through **recursive partitioning**.
- Recursive partitioning is the process of deciding if and how to split a given node into two child nodes.
- Tree splits are usually chosen to minimize the "within-node" sum of squares.
- The size of the final is determined by a process of "pruning" the tree with cross-validation determining the best place to stop pruning.

- Regression trees are an example of a more algorithmic approach to constructing predictions (as opposed to probability modeling in more traditional statistical methods) with a strong emphasis on predictive performance as measured through cross-validation.

- While single regression trees have the advantage of being directly interpretable, their prediction performance is often not that great.
- However, using collections of trees can be very effective for prediction and has been used in many popular learning methods. Examples include: random forests, boosting, and Bayesian additive regression trees (BART).
- Methods such as these can perform well on much larger datasets. We will discuss additional methods if time allows.