Regression

Prof. Eric A. Suess

March 18, 2020

Introduction

Today we will breifly discuss Regression methods and the use of Regression for Classification.

- Linear Regression/Multiple Linear Regression
- Logistic Regression
- Regression Trees
- CART

You know about Regression

Having taken a Regression class you know about

- **Linear Regression**
- **Multiple Regression**

What about. . .

- **Logistic Regression**
- **Poisson Regression**
- **Generalized Linear Models (GLMs)**

You know about Regression

The main idea with Regression is to model the relationship between a dependent variable and an independent variable(s).

To make **numeric predictions**.

Chapter 6

Read over the first half of Chapter 6, this is review.

We will try the predicting medical expenses example.

Dummy Variables

In R the lm() function that is used to fit linear regression models knows about dummy variables. There is no extra work that is need to include categorical variables into a regression model. This is because when a categorical variable is a **factor** in R, the lm() function knows the dummy variables to use.

See pages 180, 181 / 194, 195

Understanding Regression Trees and Model Trees

Last Chapter, **Trees** were used for **Classification**.

This Chapter, **Trees** are used for **Numeric Prediction**.

CART

One type of tree for prediction is **CART**, Classification and Regression Trees.

This is a bit of a misnomer, Linear Regresion methods are not used. Predictions are made based on the average value of examples that reach a leaf.

Model Trees

A second type of tree for prediction is known as **Model Trees**.

These were developed later, are less widely used but may be more powerful.

A **multiple linear regression model** is built from the exmples reaching that node.

Trees are an alternative to Regression Modeling

Trees can make predictions and can be considered as an **alternative** to regression modeling.

How are Trees built

The data are partitioned using a **divide-and-conquer** strategy according to the feature that will result in the greatest increase in **homogeneity** in the outcome after a split is performed.

For Classification Trees **entropy** is used.

For Numeric Decision Trees statistics such as **standard deviation** are used.

Example

Today we will fit a **multiple linear regression model** for the insurance data.

Example

We will look at the application of **Regression Trees** to the wine rating data.

The **rpart** package will be used.