Predicting medical expenses

Miguel Ángel Canela
Associate Professor, IESE Business School

Introduction

The objective of this project is to develop a model for predicting the **medical expenses** of the beneficiary of a **health insurance plan**. In order to make money, a health insurance company needs to collect more in yearly premiums than it spends on medical care to its beneficiaries. As a result, insurers invest a great deal of time and money in developing models that accurately predict medical expenses for the insured population.

Medical expenses are difficult to estimate because the most costly conditions are rare and seemingly random. Still, some conditions are more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart disease may be more likely among the obese.

The goal of this analysis is to use patient data to estimate the average medical care expenses for such population segments. These estimates can be used to create actuarial tables that set the price of yearly premiums higher or lower, depending on the expected treatment costs.

The data set

The data set (file `medical.csv`) contains data on 1,338 beneficiaries currently enrolled in the insurance plan, with features indicating characteristics of the patient as well as the total medical expenses charged to the plan for the calendar year.

The variables included in the data set are:

- `age`, the age of the primary beneficiary in years. There are no beneficiaries above 64 years, since they are covered by the government.
- `sex`, the primary beneficiary's gender (female/male).
- `bmi`, the primary beneficiary's **body mass index** (BMI). It is the persons weight in kilograms divided by the square of height in meters. Commonly accepted BMI ranges are: underweight (under 18.5 kg/m^2), normal weight (18.5 to 25), overweight (25 to 30), and obese (over 30).
- `dependents`, the total number of children and dependents covered by the insurance plan.
- `smoker`, a dummy indicating whether the insured regularly smokes tobacco.
- `region`, the beneficiary's place of residence in the US, divided into four geographic regions (northeast/northwest/southeast/southwest).
- `charges`, the medical costs the beneficiaries charged to the insurance plan for the year, in US dollars.

Source (slightly edited): B Lantz. Simulation based on demographic statistics from the US Census Bureau.

Questions

Q1. Use a linear regression model for predicting the medical cost in terms of the other features. How is the predictive performance of your model?

Q2. Plot the actual charges vs the predicted charges. What do you see?

Q3. We expect old age, smoking, and obesity tend to be linked to additional health issues, while additional family member dependents may result in an increase in physician visits and preventive care such as vaccinations and yearly physical exams. Is this what you find with your model? How would you modify your equation to cope better with these patterns?

Q4. Do you think that a decision tree model can work here?

Q5. Do your models overfit the data?