Background

Describe the Data Set

The data used was retrieved from [kaggle.com](http://kaggle.com) on (date). It was first published in Machine Learning with R by Brett Lantz. The data are a synthetic set intended for practice, so any conclusions we draw tell us more about the assumptions in generating the data rather than actual conditions.

The data contain the following fields:

The age of the primary beneficiary in years.

The sex of the primary beneficiary (male or female). No information is given on if this is self-identified sex or assigned based on biological characteristics of the patient.

The body mass index (BMI) of primary beneficiary, in units of kg/m2. BMI is calculated using the patient’s weight in kilograms over their height in meters squared.

The number of children on the health insurance. (While many insurance programs offer coverage for spouses and/or domestic partners, this category is labeled children, rather than ‘dependents’, so may not include spouses.)

If the primary beneficiary is a smoker.

What region of the United States the beneficiary resides in: northeast, northwest, southeast, or southwest. Regional definitions are not given in the documentation.

Costs billed by health insurance, in dollars.

Some caveats we must keep in mind. We’ll assume that intersex and transgender primary beneficiaries are a minority, so in most cases the beneficiary’s self-identified sex aligns with a typical masculine or feminine body plan. This is reasonable as (insert some data).

BMI is commonly used to make health predictions as it is far easier to calculate than more detailed measures of body composition, and on a population level it is correlated with increased body fat (check this). However, for any given individual, BMI is less reliable. In addition, ongoing study is on how well BMI serves for as a proxy for health, and to what extent it is a cause or an effect of certain health conditions. This might be academic when reporting the costs of insuring a given beneficiary, but is useful for a company looking to save money by encouraging patients to pursue preventive health care.

The region in the United States does correlate with cost of health care, but costs often change on a far more granular level, even within a large state. For a first look, a region may be enough to establish a trend; ideally a zip code is collected and data can be examined in more granularity in future surveys.

Results