Capstone Proposal

"The George Burns Effect"

How can we predict which individuals are likely to be long-lived (80+ at time of death) despite smoking throughout their entire lives? Look at data obtainable from data.gov and the CDC regarding smokers' age at death, cause of death, climate/location where they reside, race, gender, socioeconomic level and genetic markers.

Specifically, there are several pieces of information that should be obtained:

1. Self-reported smokers age 80+ (living)
2. Smokers that died age 80+ from smoking attributable causes
3. Smokers that died age 80+ from non-smoking attributable causes
4. General smoker demographics (all ages) on a national basis.

Each of the above items will include associated demographic information (race, gender, socioeconomic level and education level). Demographic information from the first three items above will be combined into a single dataframe and used to determine what types of individuals live to age 80+ despite smoking throughout their lives. (To this end, we will endeavor to omit information for long-lived individuals who quit smoking at some earlier point in their lives.) This information will then be compared with item 4, the general smoker demographics in order to make a prediction as to who is likely to survive to age 80+. (If possible, we will also attempt to compare this survival percentage to the percentage of long-lived smokers with FOXO3 and APOE Single Nucleotide Polymorphisms (SNP's) in restricted groups. This would be a one-line statement in the report only. Insufficient data exists in the proper format otherwise, nor does it appear to exist as a percentage with respect to a larger, more general population.)

The general smoking data for item 4 could be obtained in part from the following **CDC link**. It appears to be based on a survey and contains ages in broad ranges and has incomplete racial, gender and educational data (for example, "All races" and "All grades" are frequent entries). Tables 8.3 and 8.4 located **here** contain additional demographic data concerning race and gender, while additional educational/socioeconomic data is obtainable **here**. Death rates broken out by cause of death, race and gender in some detail are located in several mortality tables **here** and at the WHO **here**. Specific causes of death related to smoking could be targeted in the more general mortality tables, such as heart and lung diseases. (Those datasets currently in PDF format will be converted to a CSV file using Tabula.)

The foregoing data sources are preliminary and may be subject to changes and/or additions.

The ultimate delieverable will be in the form of a unified dataframe containing the relevant columns from the foregoing datasets along with a report containing a description of the sources, procedures followed and R code used to draw our conclusions, along with any graphs that may aid in illustrating those conclusions.

This topic would be of interest to health insurance companies in setting rates on customers who are known to be smokers and to tobacco companies who can focus their marketing efforts on those who will likely not suffer the ill effects of smoking.