Comparing the Logistic Regression Models built from the Framingham Heart Study Dataset

Problem Statement

Cardiovascular diseases (CVDs) are the leading cause of death globally, and has been for the last 15 years (WHO, 2016). In the US, 1 in every 3 deaths is attributed to CVDs (American Heart Association, 2015), and is estimated to be costing $1 billion/day in healthcare cost and lost productivity (CDC Foundation, 2015). Extensive clinical research and studies like the Framingham Heart Study have led to the identification of high blood pressure and high cholesterol as two important causal risks factors (Mahmood, Levy, Vasan, & Wang, 2014). Coronary heart disease (CHD) is the most dangerous and the most prevalent CVD. However, with proper screening and education CHD is preventable (Center for Disease Control and Prevention, 2013).

Despite decades of clinical and epidemiological research, estimating the risk of CHD remains challenging, with classification errors as high as 37% in some cases (Kones, 2011). Recently, Diabetes has been implicated as an additional risk factor, based on evidence of its comorbidity, but additional analysis could lead to further refinement of the screening programs (Fox, 2010). Clinical evidence confirms significant increases in blood glucose fluctuation in diabetic patients with CHD (compared to those without) and the important cardiovascular benefits of glycemic control agents (Cheng, Badreldin, Patel, & Bhatt, 2017; Xu & Rajaratnam, 2017). This implicates a role for blood glucose levels, as opposed to simply the presence or absence of diabetes, in the prediction model (Davidson, & Parkin, 2009; Zhang et al., 2014).

Such a model would be more accurate and improve efforts to screen for and educate about all CVDs, and particularly CHD.

Approach for the Study

This quantitative research will involve a logistic regression analysis of the Framingham Heart Study data set (that includes a series of three time measures) on the 10-year eventual outcome Coronary Heart Disease.

A Short History of the Framingham Heart Study

The leading cause of death is, and has been for the last 20 years, cardiovascular disease. Unfortunately, heart disease and stroke has been a public health epidemic for many decades. In an effort to provide a solution, the National Heart, Lung and Blood Institute (NHLBI) in collaboration with Boston University began a research project almost 70 years ago (1948) in the town of Framingham, MA. This project is known as the Framingham Heart Study and it is credited with identifying the major risks factors that contribute to CVD.

Initially, over 5,000 participants were recruited for the Framingham heart study. These men and women agreed to have certain aspects of their behavior and their results from physical and clinical test monitored over an extended period. The data collected was analyzed and revealed commonalities relating to the development of CVD. These major risk factors identified by the researchers, include high blood pressure, high blood cholesterol, smoking, age, gender and more recently diabetes. This discovery is resulted to the advancement of medical practices and to effective treatment methods for the CVD and its related diseases.

The Framingham Heart Study has since expanded to include more people and other towns, and researchers continue to collect very important data. The advancement of new medical technologies and new collaborations have improved the protocols and expansions to included

ultrasound, echocardiograph and even genetic data. Data collected are also used for projects about other diseases like dementia, diabetes, osteoporosis. There is now a Framingham Heart Study risk calculator, that allows the calculation of the probability of developing Chronic Heart disease over the next ten years of your life.

The Ten Year Chronic Heart Disease (TenYearCHD) Risk Factor calculator is available as and software application online. It requires the input of four factors: “age”, “gender”, “Cholesterol” and “Blood Pressure”. Recently, some forms of the application now include a diagnosis of “diabetes” response. The calculation of the probability of CHD is based on a model built from the Framingham Heart Study dataset. It is my contention that the measure of diabetes is not as good a risk factor as the actual measure of the blood glucose level. I have used a subset of the FHS dataset to build two models, one with and without “glucose” and evaluated the comparable quality of the two. I contend that including “glucose” specifically blood glucose levels, improve the accuracy of the model.