JOHNS HOPKINS BLOOMBERG SCHOOL OF PUBLIC HEALTH

ADVANCED DATA SCIENCE - I

Gail Score Prediction based on Facebook Profiles

Author:
Prosenjit Kundu

Supervisor:
Dr. Jeff Leek
Dr. Elizabeth Colantuoni
Dr. John Muschelli

A project submitted in fulfillment of the requirements for the Advanced Data Science - I

in the

Dept. of Biostatistics

October 18, 2016

Abstract

The main grail of this project is to estimate the ~~gail~~ score(the absolute risk of breast cancer) for a woman based on her facebook profile. Data is collected from a post in the The Breast Cancer Site page of facebook. The subjects considered are the first 2000 users who commented on the post. Based on some assumptions, covariates such as gender, race and age are predicted from names, surnames and first names of a subject using SSA information. The Gail model is implemented to estimate the absolute risk for five years and absolute lifetime risk for each of the subjects. The uncertainity in the gail score for each subject is calculated by considering the uncertainity in age and race.

Keywords : Absolute Risk, Breast Cancer, ~~BCRAT~~, Facebook, Gail Model, Median, Name-Age Calculator, Quantiles.

Acknowledgements

Contents

List of Figures

List of Tables

Chapter 1

Gail Score Prediction

1.1 Introduction

In the U.S., breast cancer is the second most common cancer in women after skin cancer. It can occur in both men and women, but it is rare in men. Breast cancer is caused due to genetic effects(mutation of BRCA1 and BRCA2) and environmental effects. Absolute risk of breast cancer or the Gail risk score is defined as the likelihood that a person who is free of breast cancer at a given age will develop that cancer over a certain period of time. An woman with a risk score of 1.4 over 5 years will be interpreted as out of a population of women with the same covariates(same age, race, etc), 1.4 % of the women will develop breast cancer in 5 years. It does not say anything about which of the ~~woman~~ in the population is going to develop breast cancer.

Women who have a Gail risk score of 1.66 or higher have a higher than average risk for developing breast cancer.

Significance: It is very important to know the risk of having a breast cancer so that the lifestyles or eating habits could be changed, appropriate medicines could be taken before the start of cancer and many other things causing cancer could be avoided accordingly.

Gail score is calculated based on the Breast Cancer Risk Assessment Tool from National Cancer Institute. The seven risk factors include age, age at first period, age at the time of the birth of her first child (or has not given birth at all), family history of breast cancer (mother, sister, or daughter), number of past breast biopsies, number of breast biopsies showing atypical hyperplasia, and race/ethnicity. In this project women with no medical history of breast cancer are considered .

Objective : Estimate the absolute risk of breast cancer for five years and lifetime absolute risk for a woman. Estimate the uncertainity in the score.

1.2 Data

Data is collected from the Facebook post. Top 2000 users who commented on that post is taken as the sample for this project. The data is read into R by creating an account in the facebook developer API and then using the package Rfacebook. ~~Access to the 241st user was denied due to privacy~~

issues. The following information for each of the users(of the 1999 users) or subjects are available from the Facebook:

- ID
- name
- first name
- last name
- picture url
- comments

Required variables for calculating the absolute risk :

- Medical history(yes or no)
- Age(35 - 90 years)
- Race

1.2.1 Processing the data

- Gender of the subject is determined from the names of the subjects using gender package in R which uses the information from the U.S. Social Security Administration baby name data. Male users are eliminated as the breast cancer for woman is considered. It worked fine with no misclassification based on a random subsample(size 25) of the given sample by just looking into the profile pics.
- Presence of medical history of breast cancer for an woman is determined by looking some of the appropriate words/sentences like "B/breast C/cancer S/survivor", "B/breast C/cancer F/free", "B/breast C/cancer D/diagnosed" in the comments assuming women commented about themselves. Women with medical history of breast cancer are removed as the gail score/absolute risk is calculated for woman free of breast cancer.
- Age is determined from the first names using the Name-Age Calculator which uses information from social security records(after 1985). The empirical distribution is seen for a particular first name and the median age is taken to be the estimated age of that woman with that first name. To get an uncertainity in the age, the 25th and 75th quantiles are taken as the bounds to form a set of probable ages. Women with first names that are not there in the records of the name age calculator are eliminated.
- Race/ethnicity is determined from the surnames assuming the surnames of the target women are same as of that before their marriage. Probabilities for five categories of ethnicity("White", "Black", "Hispanic", "Asian", "Others") are calculated using a Bayesian approach from the wru package in R by using information from the U.S. Census 2000 Surname List and Spanish Surname List. The most probable race among the five is taken as the estimated race for the woman with that

surname. Predicted races with probability < 0.01 are ignored and the rest are taken as a set of probable ages to account for uncertainity in race.

- The final data set on which the gail score and its uncertainity is estimated thus contains the names, first names, last names, median age(with a range of probable ages) and most probable race(with a range of probable races) for each woman.

1.3 Method

The probability of a woman at age α with an age-dependent relative risk $r(t)$ who will develop breast cancer by age $\alpha + \tau$ is determined by

$$P(\alpha, \tau, r) = \int_{\alpha}^{\alpha+\tau} h_1(t) r(t) e^{-\int_{\alpha}^{t} h_1(u) r(u) du} \Big\{ \frac{S_2(t)}{S_2(\alpha)} \Big\} dt \tag{1.1}$$

where h_1 is the baseline age-specific hazard of developing breast cancer, S_2 is the probability of surviving the death due to other causes, that is surviving the competing risks up to age t, S_1 is the probability of surviving the death due to breast cancer

- The absolute risk for five years and lifetime risk is calculated from the above formula where the coefficients (that is the β's) in the logistic regression are taken as the estimates estimated from a case-control data for different races .
- For a woman, a probable range of risk scores are calculated by taking different combinations of probable age and race.
- The $2.5th$ and $97.5th$ quantile of the probable risk scores are taken to get a confidence interval for the risk score.

1.4 Results

1.4.1 Tables

TABLE 1.1: Estimates of absolute risk(5 years) with confidence interval for four women.

Name	Absolute risk(5 years)	Confidence Interval
Marilyn Muller	1.11	(0.27,1.35)
Denise Gallego Moreno	0.48	(0.24, 1.07)
Sue Wedin Jones	1.03	(0.32,1.33)
Michele Byberg	0.60	(0.31,0.73)

TABLE 1.2: Estimates of lifetime risk with confidence interval for four women.

Name	Lifetime risk	Confidence Interval
Marilyn Muller	3.99	(0.72,5.36)
Denise Gallego Moreno	4.03	(1.84, 7.22)
Sue Wedin Jones	4.44	(0.99,5.85)
Michele Byberg	6.14	(3.93,6.54)

1.4.2 Figures

1.5 Data Limitations

1.6 Conclusion

1.7 References

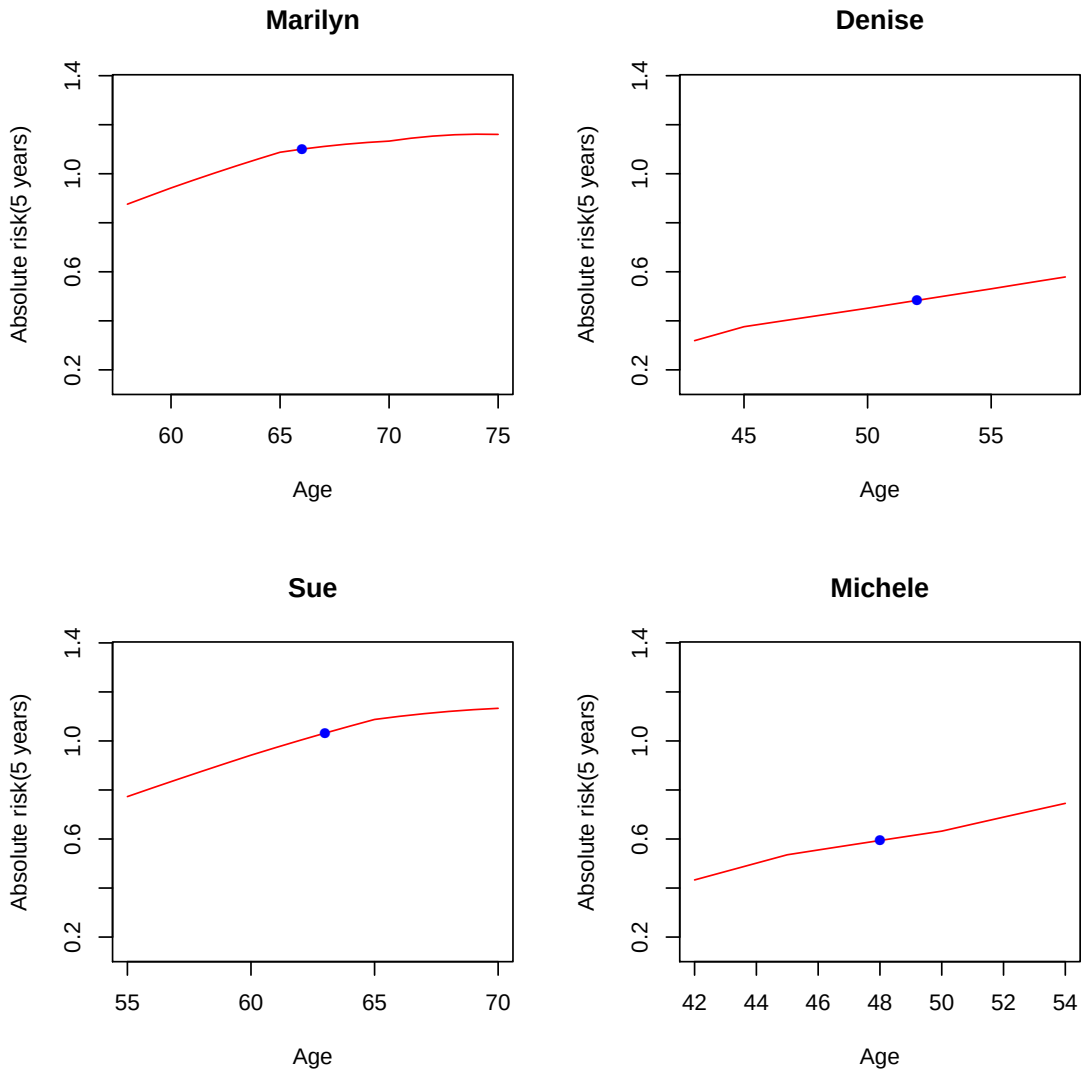

FIGURE 1.1: Absolute risk for five years versus age for four women with their race as the most probable race.

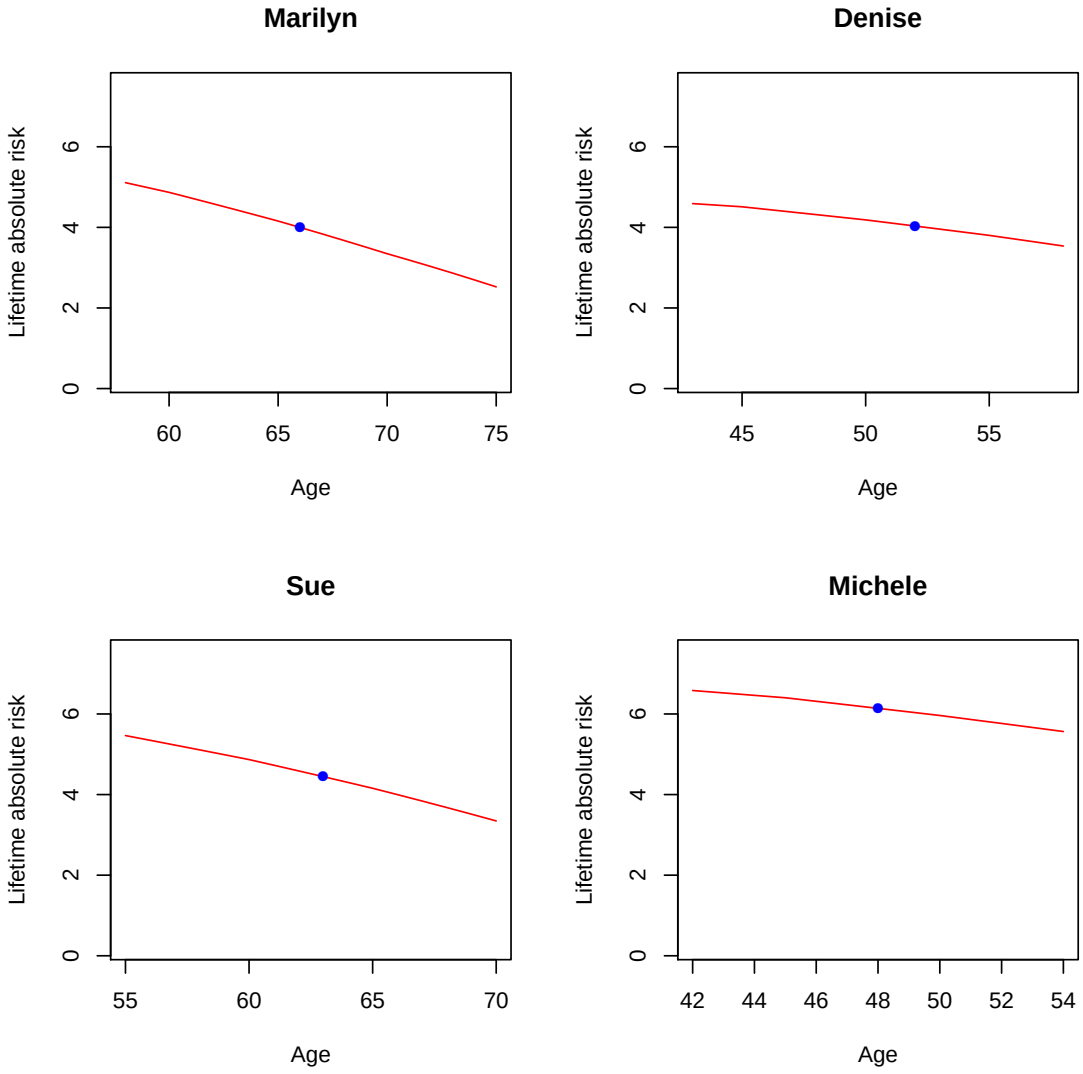

FIGURE 1.2: Lifetime absolute risk versus age for four women with their race as the most probable race.

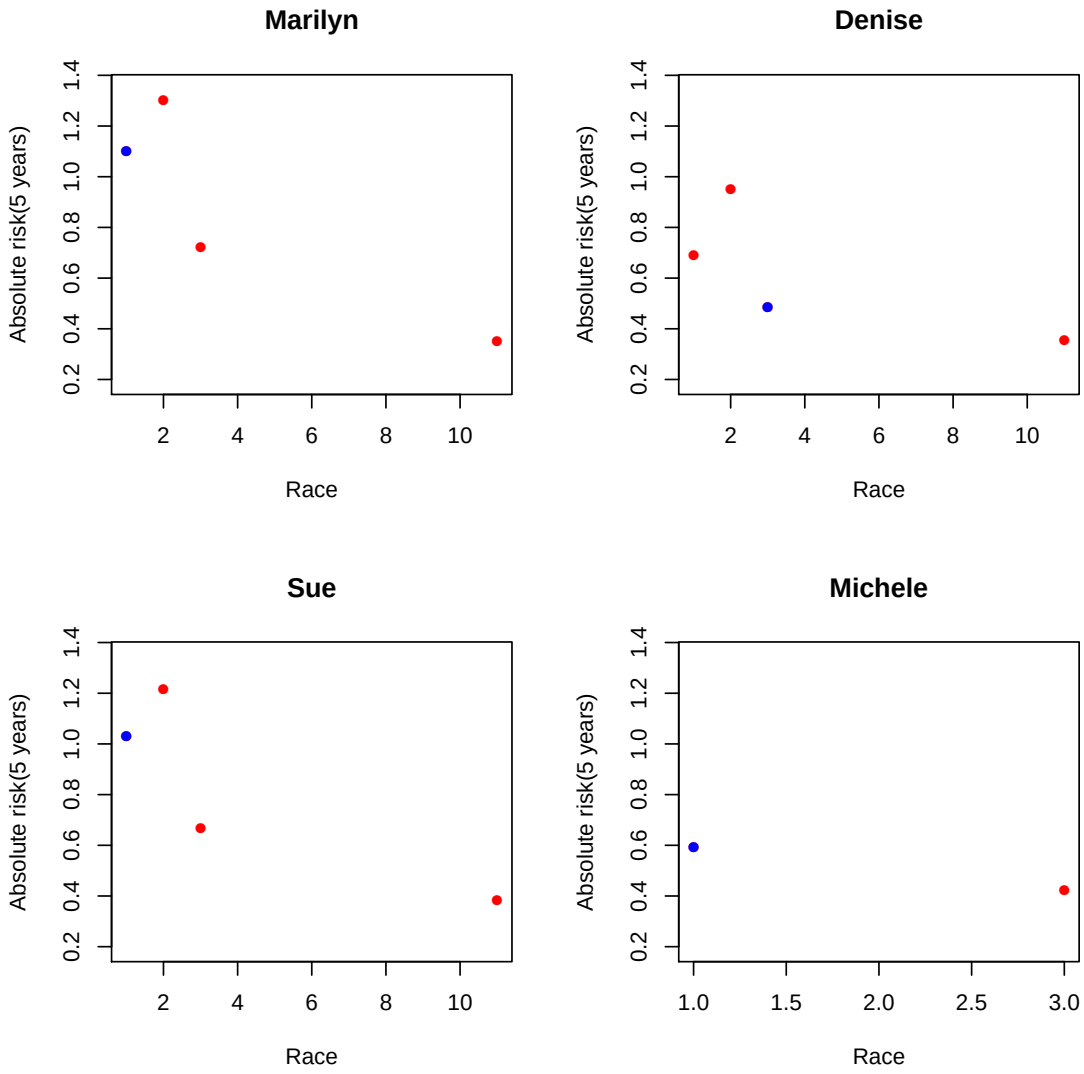

FIGURE 1.3: Absolute risk for five years versus race for four women at their median age.

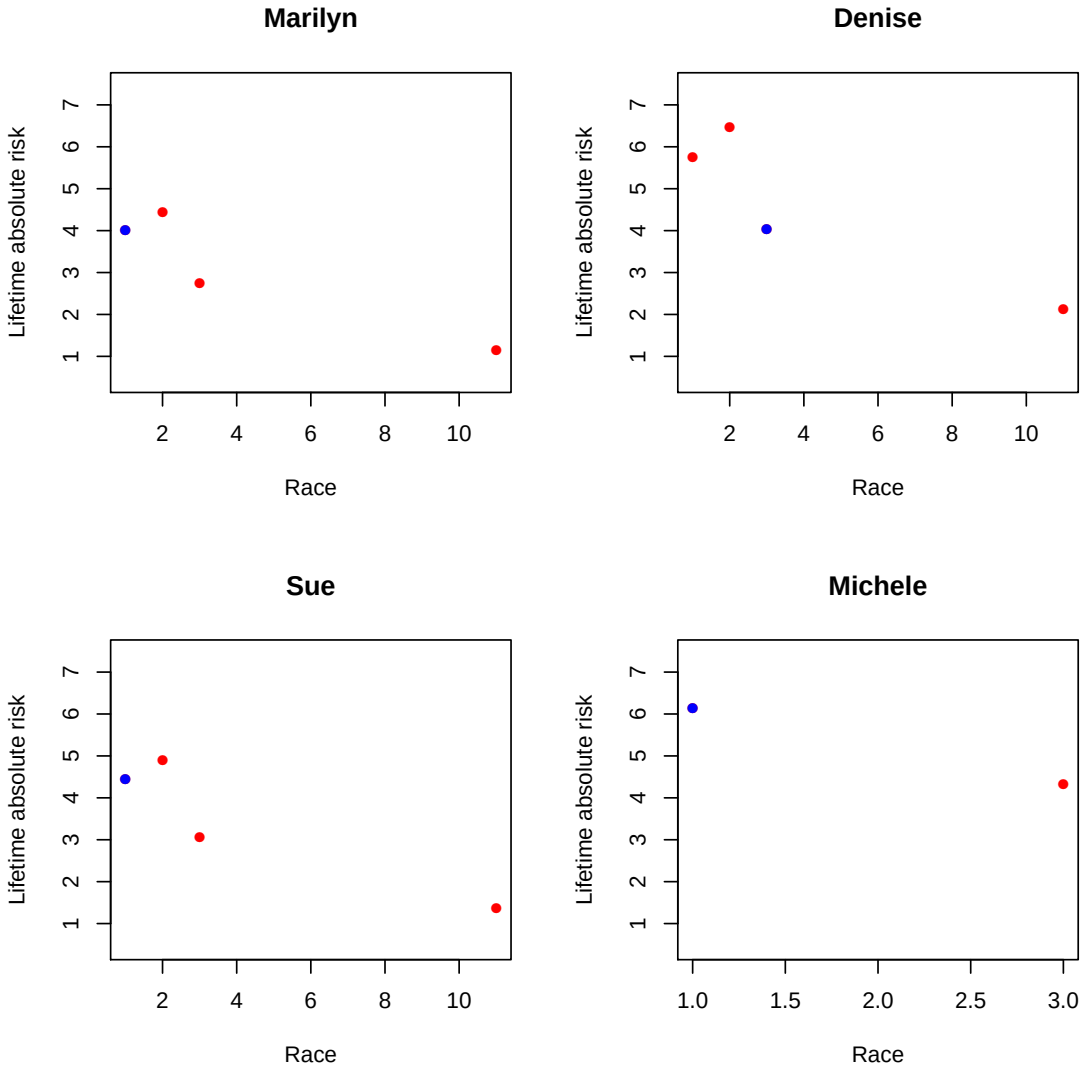

FIGURE 1.4: Lifetime absolute risk versus race for four women at their median age.

Appendix A

R Snippet

Write your Appendix content here.