Supplementary

Prosenjit Kundu

October 28, 2016

1 Checking validation of gender package

A sample of 25 subjects is taken from the post(first 25 users) and by looking at the word(him or her) of the sentence "send him/her a friend request" their gender is detected and compared with the predicted from first names using the gender package. This is shown in Table 1. NA reprsents that the name is not available in the genderdata(ssa data). If we exclude te cases, the misclassification rate is proportion of misclassifications(or Yes's) $= \frac{0}{24} \times 100\% = 0\%$

2 Checking validation of wru package

The same set of 25 users is taken and by looking into their different pictures their race is detected and compared with the predicted from surnames using the wru package. This is shown in Table 2. Here NA says that there was no picture available for that person and hence was difficult to infer on race. The misclassification rate is proportion of misclassifications(or Yes's) $= \frac{1}{23} \times 100\% = 4.34\%$.

3 More into Gail model

For calculating the relative risks the coefficients(for the seven covariates) of the logistic regression are needed and they are obtained from the fitted logistic regression for different races from the BCDDP(Breast Cancer Detection Demonstration Project data). They are provided in the Table 3 The baseline age specific hazard($h_1(t)$) of developing breast cancer is obtained from the average age-specific breast cancer rates $h_1^*(t)$ given by the following equation.

$$h_1 = h_1^* \sum_{i=1}^{l} \frac{\rho_i(t)}{r_i(t)} \tag{1}$$

where
l is the total number of risk groups, $\rho_i(t)$ is the proportion of women of age t are in the risk group i and $r_i(t)$ is the relative risk of the ith group compared to

Table 1: Comparison between gender predicted from the gender package and that by visual inspection from the facebook pictures

Name	Visual detection	Gender package	Misclassification
Naomi Downs-Cartledge	Female	Female	No
Pam Amerson	Female	Female	No
Larrie Hooper	Female	NA	NA
Cheryl Schroeder	Female	Female	No
Marilyn Muller	Female	Female	No
Athena Ziogas Splaine	Female	Female	No
Denise Gallego Moreno	Female	Female	No
Anji Worley	Female	Female	No
Jennifer Dymmel Hunter	Female	Female	No
Khadijah Gresham	Female	Female	No
Genie Stockman Harris	Female	Female	No
Patti Ravnikar	Female	Female	No
Joanna Ford	Female	Female	No
Sue Wedin Jones	Female	Female	No
Gigi Burns	Female	Female	No
Trish Speake	Female	Female	No
Mary Bolwell	Female	Female	No
Daniel Kendall	Male	Male	No
Alison Harlow	Female	Female	No
Michele Byberg	Female	Female	No
Linda Waite	Female	Female	No
Eva Melton Taylor	Female	Female	No
Wendy Jackson	Female	Female	No
Debbi Clauson	Female	Female	No
Joy Dixon	Female	Female	No

the baseline group. The age-specific hazard $h_2(t)$ of dying of causes other than breast cancer is estimated from 1979 mortality rates for all causes except breast cancer and is assumed to be same for all the subjects.

4 Extra plots

Table 2: Comparison between race predicted from the wru package and that by visual inspection from the facebook pictures

Name	Visual detection	Race from wru package	Misclassification
Naomi Downs-Cartledge	White	White	No
Pam Amerson	Female	White	No
Larrie Hooper	White	White	No
Cheryl Schroeder	White	White	No
Marilyn Muller	White	White	No
Athena Ziogas Splaine	White	White	No
Denise Gallego Moreno	Hispanic	Hispanic	No
Anji Worley	NA	White	NA
Jennifer Dymmel Hunter	NA	White	NA
Khadijah Gresham	Black	White	Yes
Genie Stockman Harris	White	White	No
Patti Ravnikar	White	White	No
Joanna Ford	White	White	No
Sue Wedin Jones	White	White	No
Gigi Burns	White	White	No
Trish Speake	White	White	No
Mary Bolwell	White	White	No
Daniel Kendall	White	White	No
Alison Harlow	White	White	No
Michele Byberg	White	White	No
Linda Waite	White	White	No
Eva Melton Taylor	White	White	No
Wendy Jackson	White	White	No
Debbi Clauson	White	White	No
Joy Dixon	White	White	No

Table 3: Coefficients in the logistic regression estimated from BCDDP data

Coefficient	White	African American	Hispanic
Intercept	0.75	-0.35	-0.75
β_1	-0.01	0.03	0.11
β_2	0.09	0.27	0.09
β_3	0.53	0.18	0.53
β_4	0.22	0.00	0.22
β_5	0.96	0.47	0.96
β_6	0.29	-0.11	-0.28
β_7	-0.19	0.00	-0.19

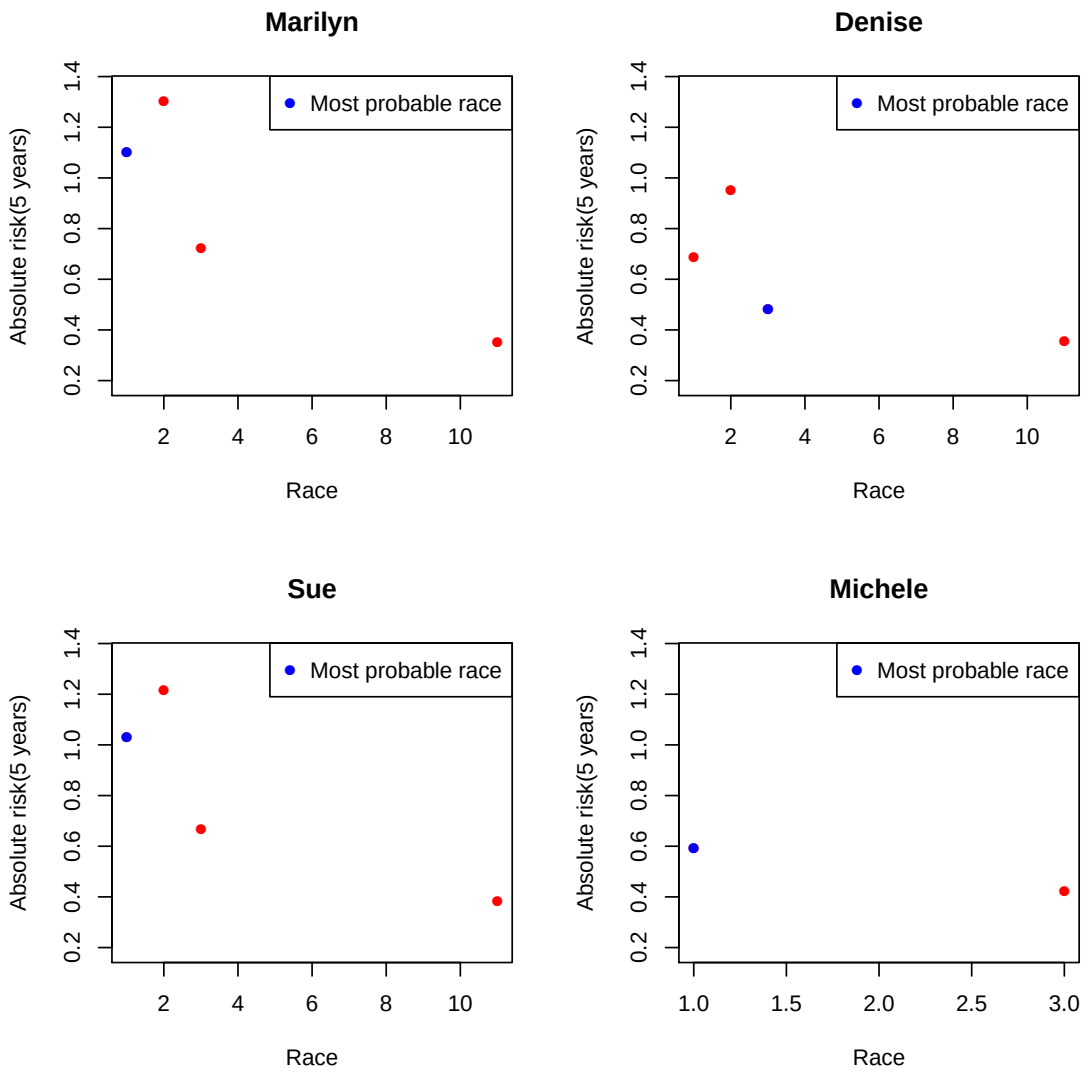

Figure 1: **Gail Score** Plot of absolute risk for five years versus race for four women with their ages fixed at their median ages. Blue dot represents the gail score at the most probable race. On the x-axis 1 represents White, 2 represents Black, 3 represents Hispanic, 11 represents Asian.

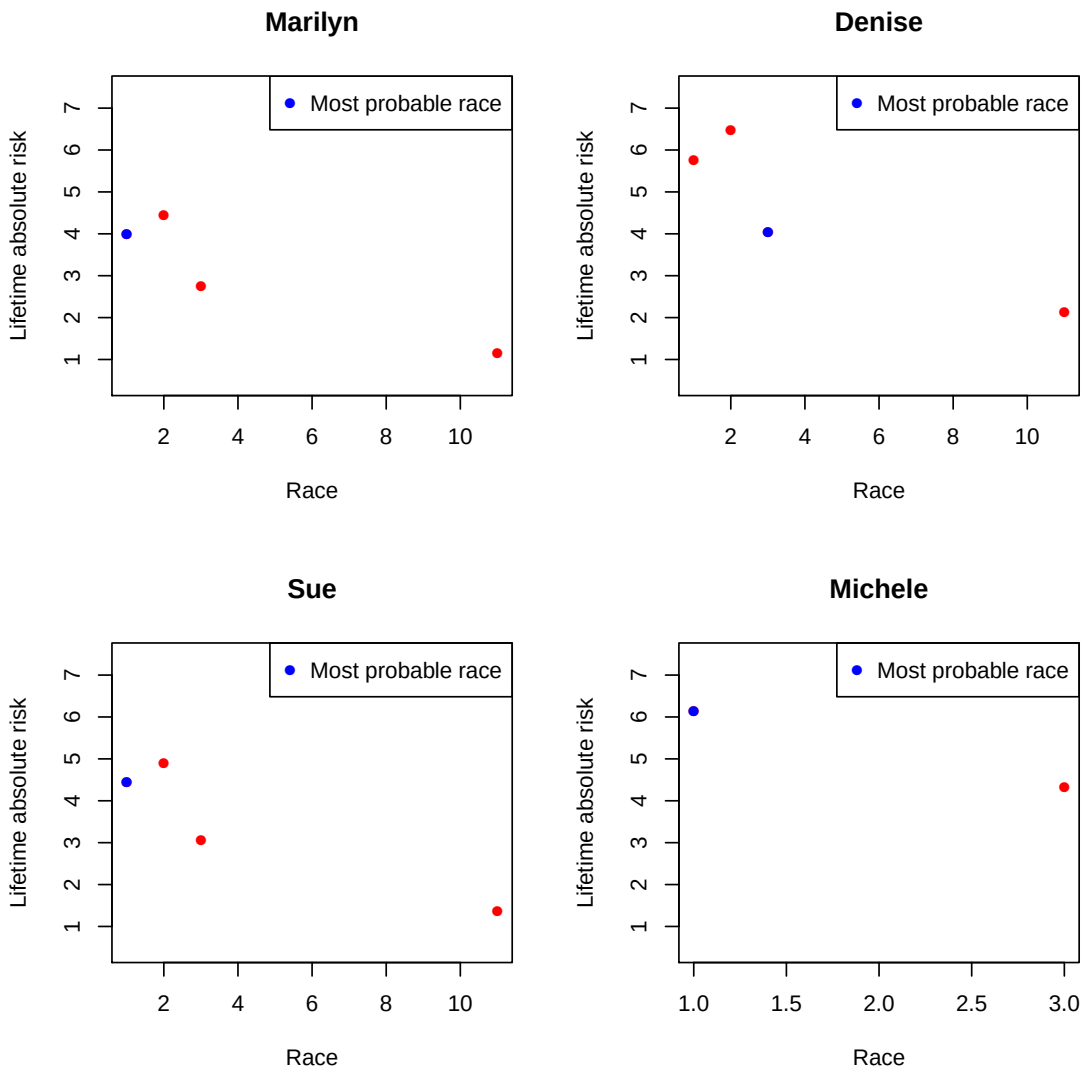

Figure 2: **Lifetime risk** Lifetime absolute risk(upto 90 years) versus race for four women with their ages fixed at their median ages. Blue dot represents the lifetime risk the most probable race. On the x-axis 1 represents White, 2 represents Black, 3 represents Hispanic, 11 represents Asian.