Biostatistics 200C: Project 1

Chad Pickering

5/11/2018

Introduction.

There is a national and global effort to increase widespread screenings for various types of cancers. This particular dataset features baseline data from the Filipino American Health Study, a randomized trial to increase colorectal cancer (CRC) screenings; the data were collected before any interventions were implemented. The study examines disparities associated with a colorectal screening test, fecal occult blood test (FOBT), versus endoscopy within the Filipino American immigrant community. Filipino Americans aged 50 and 75 from 31 community organizations in Los Angeles completed a 15-minute survey in English (65%) or Filipino (35%) about various health characteristics and economic circumstances between July 2005 and October 2006. It is of interest to understand if a patient's screening for CRC in the past is associated with all or any of the following variables: sex, age, BMI, percent lifetime spent in the U.S., annual income, having health insurance, diabetes status, and family history of cancer. We will use univariate and multivariate logistic regression to fit models that test for these potential associations.

Methods and Discussion.

The variables included in the full multivariate logistic regression model ($n = 384$ participants) are as follows. The ages of participants, in years, have a wide normal distribution with mean 59.55 and a range of 50-75. BMI ($\frac{kg}{m^2}$, derived from height and weight measures) also has a normal distribution (with one outlying point in the right tail) with mean 25.75 and standard deviation 3.7. The percentage lifetime in the U.S. is bimodal and mildly positively skewed, with median 34.41 and standard deviation 20.0. 59.4% of the participants are female, and 35.9% of participants have an income over 50K (with 5% missing). 76% of participants have health insurance, 22.9% have diabetes, and 35.2% have a family history of cancer. Sex and family history of cancer ($p = 0.018$), as well as income and having health insurance ($p < 0.001$) are correlated with each other, while all other pairs are not. Executing a logistic regression is feasible with the information obtained thus far. The regressions will not contain any transformed variables - log transforming percent lifetime in the U.S. could improve linearity slightly, but interpretation could be muddied. Additionally, using model selection procedures is not necessary here as the research question implies that all covariates need to be included in the model.

Table 3 shows the unadjusted and adjusted coefficients for the multiple logistic regression model including all eight covariates using 358 complete cases; 26 cases had missing values and were removed as advanced imputation techniques are beyond the scope of this project. 95% confidence intervals shown with each estimate were obtained using the profiled log-likelihood function. Linearity in the log-odds for the continuous variables, age, BMI, and percent lifetime in the U.S. were assessed with the Box-Tidwell test; all three had insignificant p-values suggesting that linearity assumptions hold. There appears to be no highly influential points based on raw Pearson standardized residuals, the most extreme being -2.324 (falls within any typical threshold), or changes in Pearson χ^2 values for each observation, but there are three large DFBETAs based on the $\frac{2}{\sqrt{n}} = 0.1057$ cutoff. However, upon inspection, these points do not severely violate assumptions and therefore the legitimacy of the regression estimates using all complete cases is preserved. Using the Hosmer-Lemeshow goodness-of-fit measure with $g = 10$ groups (deciles) on the full model, we find no evidence of lack of fit ($p = 0.878$) and find that it has fairly good predictive power via an ROC curve analysis (AUC=0.697). The model predicts positive and negative outcomes at about the same rate as well (positives at 64.1%, negatives at 65.4%).

There are many significant conclusions that we can draw from the multivariate logistic regression model - it suggests that the older a participant, the more of their life they have lived in the U.S., the greater their

income, and those who have a family history of cancer are significantly more likely to have had a CRC screening in the past. More specifically, for each additional year of age of a Filipino immigrant, the odds of having had a CRC screening in the past is multiplied by a factor of between 1.011 and 1.095, adjusting for all of the other covariates ($p = 0.012$). Similarly, when the percentage of a Filipino immigrant's lifetime spent in the U.S. increases by 1 percentage unit, the odds of having had a CRC screening is multiplied by a factor of between 1.012 and 1.038, adjusting for the other predictors ($p < 0.001$). Additionally, the odds of having had a CRC screening is multiplied by a factor of between 1.033 and 2.974 for those who have an income over 50,000 dollars compared to those who do not, adjusting for the other covariates ($p = 0.038$). Finally, the odds of having had a CRC screening is multiplied by a factor of between 1.030 and 2.664 for those who have family history of cancer compared to those who do not, adjusting for the other predictors ($p = 0.038$). All ranges given here are 95% confidence intervals (Table 3).

When adjustments for other variables are not made, e.g. odds ratio estimates are obtained from separate univariate logistic models, the qualitative conclusions hold regarding age ($p = 0.043$), percentage of lifetime spent in the U.S. ($p < 0.001$), and income ($p = 0.001$) (Table 2). When not adjusted for, having insurance increases the odds of having had a CRC screening by a multiplicative factor of between 1.698 and 4.711 compared to those who do not have insurance ($p < 0.001$). This significance is a huge shift from its impact in the multivariate model ($p = 0.381$), probably due to its collinearity with income. Lastly, both diabetes and family history of cancer are borderline significant (both $p = 0.053$) when the other covariates are not included in their respective models.

Limitations and conclusion.

The 8-variable logistic regression model used throughout was probably not the best predictive model, but excluding any of the variables would remove some of the vehicles of adjustment required by the research question. One would want the best predictive model if targeting new screening patients in the Filipino American immigrant population is desired. This model could also account for interactions between predictors, a feature not explored here. The survey itself could also be edited to include new questions inquiring about other diseases and medical conditions, and the survey design could be expanded to include younger participants.

In conclusion, having had a CRC screening is associated with age, time spent living in the U.S., income level, and a positive family history of cancer, and potentially weakly associated with owning health insurance and having diabetes.

Tables.

Table 1a. Summary statistics of continuous variables.

Variable	Mean (Median)	St. Dev.	Missing
Age (years)	59.55 (59.0)	6.2	7 (0.018)
BMI ($\frac{kg}{m^2}$)	25.75 (25.3)	3.7	1 (0.003)
Pct. lifetime in U.S.	33.25 (34.4)	20.0	8 (0.021)

Table 1b. Summary statistics of non-continuous variables.

Variable	Yes/1	No/0	Missing
Sex (1=male)	156 (0.406)	228 (0.594)	0 (0.0)
Annual income, >50K (dollars)	138 (0.359)	227 (0.591)	19 (0.050)
Has health insurance	292 (0.760)	92 (0.240)	0 (0.0)
Has diabetes	88 (0.229)	296 (0.771)	0 (0.0)
Has family history of cancer	135 (0.352)	249 (0.648)	0 (0.0)

Table 2. Odds ratios for univariate (unadjusted) models.

Term	OR (95% CI)	P-value
Age	1.035 (1.001, 1.070)	**0.043**
Sex (male)	1.313 (0.873, 1.978)	0.191
BMI	1.020 (0.966, 1.078)	0.469
Pct. lifetime	1.031 (1.020, 1.043)	**<0.001**
Income	2.024 (1.320, 3.120)	**0.001**
Insurance	2.794 (1.698, 4.711)	**<0.001**
Diabetes	1.604 (0.995, 2.601)	0.053
Fam. history	1.517 (0.996, 2.315)	0.053

Table 3. Multiple logistic regression model (n=358).

Term	Unadj. Estimate (95% CI)	P-value	Adj. Estimate (95% CI)	OR (95% CI)	P-value
Intercept			-4.089 (-7.046, -1.186)		**0.006**
Age	0.034 (0.001, 0.068)	**0.043**	0.051 (0.011, 0.091)	1.052 (1.011, 1.095)	**0.012**
Sex (male)	0.273 (-0.136, 0.682)	0.191	0.132 (-0.335, 0.559)	1.141 (0.715, 1.821)	0.580
BMI	0.020 (-0.034, 0.075)	0.469	-0.022 (-0.085, 0.039)	0.978 (0.918, 1.040)	0.485
Pct. lifetime	0.030 (0.019, 0.042)	**<0.001**	0.024 (0.012, 0.037)	1.024 (1.012, 1.038)	**<0.001**
Income	0.705 (0.278, 1.138)	**0.001**	0.558 (0.033, 1.090)	1.747 (1.033, 2.974)	**0.038**
Insurance	1.028 (0.529, 1.550)	**<0.001**	0.268 (-0.330, 0.875)	1.308 (0.719, 2.399)	0.381
Diabetes	0.473 (-0.005, 0.956)	0.053	0.371 (-0.157, 0.905)	1.450 (0.855, 2.471)	0.169
Fam. history	0.417 (-0.004, 0.839)	0.053	0.502 (0.029, 0.980)	1.652 (1.030, 2.664)	**0.038**