Project - Part #2
Due Thursday, November 29, 2007

The data are subsets from the Project 1 data.

1. We first consider a very small subset of the data for women aged 45-54 who do not have missing values on any of the covariates. The records for 3,195 women can be found in the dataset *risk2.dta* on the website. We have included all the breast cancer cases and randomly sampled 1% of the controls. The scientific interest is in breast density as a *grouped linear* variable. We want to investigate the relationship of breast cancer to increasing breast density. We consider other possible confounders: age (2 levels); family history (2 levels); BMI (4 levels treat as categorical); menopausal status (2 levels); and hormone replacement therapy (2 levels). Note that HRT is coded as zero for premenopausal women.

    a. Write the theoretical model for the logit for a diagnosis of breast cancer if it depends on breast density, age, BMI, menopausal status, and HRT use. You can assume no interactions.

    b. Since we sampled controls, but not cases, which of these coefficients would be most affected? Would it change the odds ratios or their precision?

    c. Fit a model in Stata that includes breast density only as a grouped linear variable.

    d. Fit a model in Stata that includes breast density as grouped linear variable and adjusts for all potential confounders. Compare the odds ratios for breast density of this model to (c) and decide if the other variables are “confounders”.

    e. Compute the odds ratio for a post-menopausal woman using HRT to that of a premenopausal woman of the same age, same BMI, same breast density, and same family history.

    f. Test whether BMI is an effect modifier of the relationship of breast density to breast cancer outcome.

    g. Create a binary variable representing high breast density (coded as 1 if BIRADS breast density is 3 or 4; coded as 0 if BIRADS breast density <=2). Now do a propensity score analysis by using the covariates of (d) (excluding density and cancer status) to predict high breast density. Interpret *scientifically* how these covariates relate to high breast density. Keep all covariates regardless of significance and estimate fitted probabilities and the linear predictor.

    h. Stratify by the propensity to have high breast density and fit a conditional logistic regression model of case status and breast density. Compare the OR to the model of (d) above.

i. Using the same model as in (d), add the linear predictor from (g) to the standard logistic regression model. What is the effect on the OR for breast density?

j. Using the same model as in (c), add the linear predictor from (g) to the standard logistic regression model. What is the effect on the OR for breast density?

2. For postmenopausal women we are interested in breast density as a grouped linear variable and its association with invasive breast cancer. We eliminate all women with unknown values for breast density, HRT, BMI, family history (coded as negative (nrelbc=0) or positive (nrelbc=1 or 2). We want to compare the cohort analysis using all the data to a matched case-control study.

   a. Using the data set from Project 1 get the OR and 95% CI for breast density adjusting for the main effects of age (categorical), HRT, BMI (grouped linear), family history. Be sure to eliminate unknown values and recode family history to be a dichotomous variable.

   b. What what would be the theoretical model for the log odds ratio in part (a)?

   c. We then do 1-1 matching of cases to controls by age (age.dta). We use all the cases, but randomly select a control from the same age group for each case. What parameters in (b) would be eliminated in this conditional model?

   d. Using the data matched on age, determine the odds ratio for breast density from the matched sets. Repeat grouping on age, rather than the actual matched sets. Compare these two OR's for breast density and their CI's.

   e. Using the data matched on age, determine the odds ratio for breast density from the matched sets, but adjusting for HRT, family history, and BMI (linear) as potential confounders. Which of these other variables appear to be significant predictors as well? Compare the OR and CI for breast density in this model to that of (a).

   f. Determine if any of the variables in (e) are also effect modifiers of the density OR. Also test age as an effect modifier, but only as a linear x linear term (e.g. age*density).

   g. We also matched on (1) age and HRT; (2) age and BMI; and (3) age, BMI, and HRT. We will consider smaller studies more typical of case-control studies. Use only one dataset (see assigned dataset on last page). For your dataset, what odds ratios are estimable? Compare the odds ratio for density in your dataset unadjusted for other variables versus adjusted for main effects of variables that can be included. Also compare the adjusted estimate to the OR for density in (e). In discussion section we will compare the OR's and CI's for density adjusted for all main effects possible.

Risk Model Dataset Documentation

Biost 536 Project Part 2

Question 1 (restricted to women aged 45-54)
In file risk2.dta

Variable	Coding
menopaus	0 = premenopausal; 1 = postmenopausal or age>=55
age50	0 = 45-49; 1 = 50-54
density	BI-RADS breast density codes 1 = Almost entirely fat; 2 = Scattered fibroglandular densities; 3 = Heterogeneously dense; 4 = Extremely dense
bmi	Body mass index: 1 = 10-24.99; 2 = 25-29.99; 3 = 30-34.99; 4 = 35 or more
famhx	Family history of breast cancer: 0 = No; 1= Yes
hrt	Current hormone therapy: 0 = no; 1 = yes; coded as 0 for pre-menopausal women
invasive	Diagnosis of invasive breast cancer within one year of the index screening mammogram: 0 = no; 1 = yes

Question 2 (restricted to post menopausal women with complete covariate information; controls randomly sampled for matched sets)

In file age.dta (matched on age), agehrt.dta (matched on age and hrt), agebmi.dta (matched on age and bmi), agebmihrt.dta (matched on age, bmi, and hrt)

Variable	Coding
set	Matched set ID
agegrp	3= 45-49; 4 = 50-54; etc
density	BI-RADS breast density codes 1 = Almost entirely fat; 2 = Scattered fibroglandular densities; 3 = Heterogeneously dense; 4 = Extremely dense
bmi	Body mass index: 1 = 10-24.99; 2 = 25-29.99; 3 = 30-34.99; 4 = 35 or more
famhx	Family history of breast cancer: 0 = No; 1= Yes
hrt	Current hormone therapy: 0 = no; 1 = yes
invasive	Diagnosis of invasive breast cancer within one year of the index screening mammogram: 0 = no; 1 = yes

Assignment for problem 2g: (You will work individually, but may have the same dataset).
These have changed since they were first posted. If you already did the original assignment just turn that one in.

Last name	Assigned dataset
Bolton, Dean, Devine	use agehrt.dta if set<=541
Doherty, Don	use agehrt.dta if set>541&set<=1083
Gant, Hardikar, Heike	use agehrt.dta if set>1083&set<=1624
Horne, Hsia	use agehrt.dta if set>1624
John, Katz, Kim	use agebmihrt.dta if set<=541
Kross, Lambdin	use agebmihrt.dta if set>541&set<=1083
Liu, Livaudais, Nelson	use agebmihrt.dta if set>1083&set<=1624
Pattrapornnan, Rees, Yu	use agebmihrt.dta if set>1624
Ringold, Rodriguez, Sakoda	use agebmi.dta if set<=541
Scott, Shah, Zhou	use agebmi.dta if set>541&set<=1083
Sugimoto, Upson, Verrall	use agebmi.dta if set>1083&set<=1624
Waweru, Wong, Zhao	use agebmi.dta if set>1624