Biostatistics 140.654
Fourth Term, 2021
Problem Set 1

Instructions: Students may work on this assignment in groups/teams; however, required written summaries should be completed alone. You may serve as editors for other students; i.e. you may ask a friend and vice versa to help review your written summaries.

Due date: Friday, April 9th by 5pm EST

I. Rudimentary understanding of logistic regression analysis

Upon mastery of this problem, a student should be able to:

- teach logistic regression for binary responses to a health scientist or professional; interpret regression coefficients as log odds ratios
- graphically display binary data and make sensible guestimates of the coefficients that would be obtained from a simple logistic regression (SLR)

Simulate a data set of 250 observations that satisfy each of logistic models A and B below in which log odds(Y=1) = $\beta_0 + \beta_1 X$, for X_i iid uniform(0,1):

Model A: $\beta_0 = -2$; $\beta_1 = 2$
Model B: $\beta_0 = -2$; $\beta_1 = 4$

For each data set, fill in the table below by completing the following:

1. plot Y against X, jittering the Ys so you can see all of the observations
2. split the data set into 5 roughly equal-sized X strata by using cut points: 0.2, 0.4, 0.6, 0.8
3. estimate the log odds(Y=1) in each stratum in the table provided below
4. plot the log odds(Y=1) against the mid-point of each X stratum
5. estimate the intercept and slope for the plot of log odds(Y=1) against stratum midpoints using the graph.
6. determine the predicted log odds at the midpoint of each bin from the fitted line
7. calculate the corresponding predicted probability that Y=1 given X as a function of X.

Model	Stratum of X	N	# y=1	p= y/n	odds= p/(1-p)	$\log_e$ odds	predicted log odds	predicted Pr(Y=1)
A	0.0-0.2							
	0.2-0.4							
	0.4-0.6							
	0.6-0.8							
	0.8-1.0							
B	0.0-0.2							
	0.2-0.4							
	0.4-0.6							
	0.6-0.8							
	0.8-1.0							

You have successfully conducted two simple logistic regressions by hand.

Write a short paragraph that explains logistic regression to a layperson in your own words.

II. Connection of logistic regression to 2x2 tables; confounding and effect modification

Upon mastery of this problem, a student should be able to:

- create one or multiple 2x2 tables from which to estimate log odds ratios that correspond to coefficients from simple or multiple logistic regressions
- appreciate the invariance of the odds ratio as one important reason logistic regression is popular in epidemiology
- pool log odds ratios across strata using weighted averages as an approximation to logistic regression

Use the National Medical Expenditure Survey (NMES) data set for this problem. The general goal is to describe the association of self-reported smoking with the indicator of major smoking-caused disease (mscd), a group of diseases the U.S. Surgeon General and WHO say are caused by smoking.

Part A: Simple logistic regression

1. Define a variable mscd to represent whether or not a person has a major smoking caused disease (e.g. lc5 or chd5 =1). Make a 2x2 table of *mscd* against *eversmk* (1-yes; 0-no). Calculate the log odds ratio, its standard error and 95% CI using 652 methods for 2x2 tables. To simplify the analysis, drop those people who have a missing value of *eversmk* (this is to simplify the exercise and is not generally an acceptable strategy).

2. Regress *mscd* on *eversmk* using logistic regression. Compare the regression coefficient and its standard error with the log odds ratio and standard error in Part A Question 1 above.

3. Use logistic regression to regress *eversmk* on *mscd*. Compare the log odds ratio and standard error from this regression with those from Part A Questions 1 and 2.

4. Write a couple of sentences that can be used to teach a public health professional: the interpretation of the logistic regression coefficient; and the invariance principle of the odds ratio.

5. *Extra enjoyment.* Review the paper by Prentice and Pyke (*Biometrika*, 1979) and then state the invariance property of the log odds ratio estimate from a logistic regression in precise mathematical terms.

Part B. Association of *eversmk* and *mscd*, controlling for age.

1. Stratify age by: <50, 51-60, 61-70, >70. Within each stratum, calculate the log odds ratio and standard error for the *mscd-eversmk* association. Complete the table below. Here, *weight* is defined to be the inverse of the variance normalized to sum to 1.0 across strata: $weight_j=(1/se_j\hat{}2)/sum_j(1/se_j\hat{}2)$

age stratum (j)	log odds ratio	std error (se_j)	$1/(se_j)^2$	weight $(1/se_j^2)/sum_j(1/se_j^2)$
<50				
51-60				
61-70				
>70				
Ignoring age				

2. Calculate the weighted average log odds ratio from the data above and its standard error (= sqrt{$1/sum_j[1/se^2_j]$ }). Compare this value to the one from Part 1. Question 1 where age was not controlled by plotting each coefficient with its confidence interval on the same set of axes. Add any additional relevant information for evaluating whether age is a confounder to your figure. Explain in a sentence or two whether age is a "confounder" of the smoking-disease association and why?

3. Use logistic regression to regress *mscd* on *eversmk* and 3 indicator variables for the 4 age strata, i.e. make age category a factor. Compare the resulting *eversmk* coefficient and standard error with the value above in Part 2 Question 2.

4. Now repeat the analysis controlling for age with your favorite smooth function of continuous age with three degrees of freedom.

5. Use logistic regression to regress *eversmk* on *mscd* and the same function of age. Compare the *mscd* coefficient and standard error from this model with the *eversmk* coefficient from the model in Part 2 Question 4.

6. Plot the *mscd* data against age using the *eversmk* value as the plotting symbol or color. Add the predicted values from your model and a kernel smoother fit separately to each smoking group for comparison. Compare the model predictions with the kernel smoothers to see if there is evidence of effect modification of the smoking-mscd association by age?

7. Propose an extended model to directly address the possibility that age modifies the effect of smoking on disease prevalence. Fit this model and compare it to the model without effect modification using a likelihood ratio test.

III: Now you are the course instructor!

Using the data and analyses you did in this homework as an example, prepare an Rmd file and accompanying vignette (written or audio and/or video) that can be used to teach logistic regression to a public health professional. You can work in groups on the vignette; give credit to everyone in your group in your submission (i.e. give the author list). Be numerate, avoid non-essential statistical jargon and be as clear as possible. Remember to emphasize: Question, Question, Question.