Biostat/Epi 536 2024
HW 7.

This assignment has 11 questions. You should do all 11 questions. However, you should only submit answers to questions 3, 4, 5, 7, 8, 9, and 11. These questions are marked with ** below.

Trichopoulos et al. (1976, British J. Obstet Gynaecol 83:645) reported on a matched case-control study of secondary infertility. They were interested in spontaneous abortion and induced abortion as risk factors. The data were re-analyzed by Hogue (1978). Both papers are available in a single PDF file on Canvas folder. You are **not** required to read the papers, but they are provided if you would like to read them. **A short presentation introducing HW7 will be given in class on November 19;** this presentation should cover the necessary background from the papers.

Although you do not need the data to do this assignment, the data are available in infert.dta . Note that “the number of previous spontaneous abortions” and “the number of previous induced abortions” are recorded as in Table 1 of the paper by Trichopoulos et al. In addition to the variables listed in the paper, dummy variables g2-g8 are in the dataset to indicate all but the 0/0 and 2+/2+ categories of Table 1. A variable matchset has also been created to indicate the 83 matched sets. Moreover, strata have been formed by pooling all matched sets with identical values of the matching variables. These strata are identified by the variable called stratum. The single person who had both 2+ spontaneous abortion and 2+ induced abortions is omitted from the dataset.

Files containing software output (both R and STATA) from fitting a variety of models and are also available on CANVAS. Notice that the R output gives model coefficients and the STATA output gives Odds Ratios. Examine this output to answer the following questions.

1. Models using **ordinary** logistic regression and no adjustment for the matching variables:
(a) Model 1a) is a logistic model with only dummy variables indicating the three groups of number of prior induced abortions. According to this model, is induced abortion associated with secondary infertility?
(b) Model 1b) is a logistic model with only the dummy variables indicating the three groups of number of prior spontaneous abortions. According to this model, is spontaneous abortion associated with secondary infertility?
(c) Model 1c) is a logistic model with main-effect dummy variables for both spontaneous abortion and induced abortion. According to this model are spontaneous abortion and induced abortion associated with the risk of secondary infertility when each is adjusted for the other?
(d) Model 1d) is a model with seven dummy variables indicating the eight combinations of the spontaneous and induced abortion variables that are observed in these data. Observe the LR (likelihood ratio) test comparing this model to the model fit in c) to test for interaction between number of spontaneous abortions and number of induced abortions. What do you conclude?

In order to help you get started on the homework, here is the answer to Q1a: “According to this model, we do not have evidence that induced abortion is associated with secondary infertility (p-value = 0.96).” As you can see, the question is just asking you to identify the relevant part of the output and the appropriate P-value. You should follow that same approach for similar questions.

2. Models 2a) through 2d) repeat 1a) through 1d), also using **ordinary** logistic regression, but adjusting for matching variables as follows:
i. education: dummy variable main effects only

ii. age: continuous linear main effect only
iii. parity: dummy variable main effects only
For each of a) through d) compare results to those in Question 1 and discuss reasons
for the similarities and differences you see.

3. Models 3a) through 3d) repeat models 1a) through 1d) using a **conditional** logistic regression for which each matched set is a single stratum. For each of a) through d) compare model results to those in a) through d) of question 2, and discuss reasons for similarities and differences you see.

4. Models 4a) through 4d) repeat models 1a) through 1d) using **ordinary** logistic regression and dummy variable adjustment for each matched set. For each of a) through d) compare model results to those in 2 and 3 above, and discuss reasons for the similarities and differences you see.

5. Models 5a) through 5d) repeat models 1a) through 1d) using **conditional** logistic regression using the strata formed by pooling matched sets with like values of the matching variables. For each of a) through d) compare model results to those in question 3 and discuss reasons for the similarities and differences you see.

6. Models 6a) through 6d) repeat models 1a) through 1d) using **ordinary** logistic regression with dummy variable adjustment for the strata formed by pooling matched sets with like values of the matching variables. For each of a) through d), compare results to those in 5 above. Discuss reasons for the similarities and differences you see.

7. For which of the 24 models examined in this homework do the odds ratio estimates most closely resemble those given by Trichopoulos et al., in Table 1? That is, under which model are their estimates computed? What does this say, by today's standards, about the appropriateness of the analysis they performed? What are its strengths and weaknesses?

8. If the 24 models examined in this homework had dichotomized the two exposure variables as Hogue did, which models do you think would have led to estimates the most similar to the ones Hogue presented in both Tables 41-1 and 41-2? That is, which models do you think are closest to those under which Hogue's estimates are computed? What does this say, by today's standards, about the appropriateness of the analyses she performed? What are their strengths and weaknesses?

9. Using your answers to Questions 7 and 8, but also any comparisons in questions about these data from the other parts of this Homework, how appropriate are the estimates presented by Trichopoulos et al. in Table 1 and by Hogue in Tables 41-1 and 41-2?

10. If you were to choose one of the 24 models presented in this Homework to summarize these data in a paper on the relationship of induced abortion to the risk of secondary infertility, which would you choose, and why?

11. Discuss strengths and limitations of the study design, focusing on the investigator's decision to use a matched design, individual vs. frequency matching, the choice of matching variables, etc.