BIOST 536: Homework 7

Department of Biostatistics @ University of Washington

25 November 2024

Background

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. Ordinary unadjusted logistic regression

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?

According to this model, we do not have evidence induced abortion is associated with secondary infertility (Wald p=1.0).

(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?

According to this model, we have evidence induced abortion is associated with secondary infertility (Wald p<0.001).

(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?

According to this model, when adjusting for each other, we have insufficient evidence induced abortion is associated with secondary infertility (Wald p=0.1) and sufficient evidence spontaneous abortion is associated with secondary infertility (Wald p<0.001).

(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?

According to this model, we have insufficient evidence the number of induced abortions modify the association of number of spontaneous abortions with secondary infertility, or vice versa (LRT p=0.097).

2. Ordinary logistic regression adjusted for covariates

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.

According to the (2a) model adjusting for education, age, and parity, we do not have evidence induced abortion is associated with secondary infertility (Wald p=1.0); this agrees with the unadjusted model. Adjustment of these variables did not impact the association of secondary infertility with number of induced abortions because this association is measured within our matched-case sample, such that those with and without secondary infertility have the same distributions of age, education level, and parity.

According to the (2b) model adjusting for education, age, and parity, we have evidence spontaneous abortion is associated with secondary infertility (Wald p<0.001); this agrees with the unadjusted model. Adjustment of these variables did not impact the association of secondary infertility with number of spontaneous abortions for reason similar to those explained in 2a.

According to the (2c) model, when adjusting for education, age, parity, and each other, we have evidence induced abortion is associated with secondary infertility (Wald p<0.001) and evidence spontaneous abortion is associated with secondary infertility (Wald p<0.001). This second model disagrees with its predecessor regarding the adjusted "effect" of number of induced abortions. Compared to the (1c) model, the association of induced abortions is much stronger in the (2c) model.

According to the (2d) model, we have insufficient evidence the number of induced abortions modify the adjusted association of number of spontaneous abortions with secondary infertility, or vice versa (LRT p=0.5); this agrees with (1d).

3.* Conditional logistic regression for matched sets

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.

According to the (3a) model, we do not have sufficient evidence induced abortion is associated with secondary infertility (Wald p=0.9); this agrees with (2a).

According to the (3b) model, we have sufficient evidence spontaneous abortion is associated with secondary infertility (Wald p<0.001); this agrees with (2b).

According to the (3c) model, when adjusting for each other, we have evidence induced abortion is associated with secondary infertility (Wald p<0.001) and evidence spontaneous abortion is associated with secondary infertility (Wald p<0.001); these results agree with (2c).

According to the (3d) model, we have insufficient evidence the number of induced abortions modify the adjusted association of number of spontaneous abortions with secondary infertility, or vice versa (LRT p=0.39); this agrees with (2d). All the OR estimates of (3d) are greater than (2d)–some are enormous.

4.* Ordinary logistic regression adjusted for matched sets

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.

According to the (4a) model, we do not have sufficient evidence induced abortion is associated with secondary infertility (Wald p=0.9); this agrees with (3a, 2a, 1a).

According to the (4b) model, we have sufficient evidence spontaneous abortion is associated with secondary infertility (Wald p<0.001); this agrees with (3b, 2b, 1b).

According to the (4c) model, when adjusting for each other, we have evidence induced abortion is associated with secondary infertility (Wald p<0.001) and evidence spontaneous abortion is associated with secondary infertility (Wald p<0.001); these results agree with (3c, 2c) and disagree with (1c) regarding the significance of association with induced abortion.

According to the (4d) model, we have sufficient evidence the number of induced abortions modify the adjusted association of number of spontaneous abortions with secondary infertility, or vice versa (LRT p=0.18); this disagrees with (3d, 2d). Similar to before, some OR estimates are enormous.

5.* Conditional logistic regression for pooled matched sets

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.

According to the (5a) model, we have insufficient evidence induced abortion is associated with secondary infertility (Wald p=0.9); this agrees with similar previous models.

According to the (5b) model, we have sufficient evidence spontaneous abortion is associated with secondary infertility (Wald p<0.001); this agrees with similar previous models.

According to the (5c) model, when adjusting for each other, we have sufficient evidence induced abortion is associated with secondary infertility (Wald p<0.001) and sufficient evidence spontaneous abortion is associated with secondary infertility (Wald p<0.001); this agrees with all similar previous models except (1d).

According to the (5d) model, we have insufficient evidence the number of induced abortions modify the association of number of spontaneous abortions with secondary infertility, or vice versa (LRT p=0.3); this agrees with (3d, 2d) and disagrees with (4d). Again, notice that some OR estimates are massive.

7.* Identifying the Trichopoulos model

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?

The odds ratio estimates presented by Trichopoulos et al. closely resemble Model 1d, the logistic regression model that includes interaction between the number of induced and spontaneous abortions, without adjustment for any covariates. The absence of adjustment variables in their model is a clear limitation, as their estimates are ignorant to possible confounders. Fortunately, they discuss this weakness and present their estimates as associative. Arguably, model simplicity is a strength, but I believe the setting of their work would find an adjusted model to be more scientific and take advantage of available information.

8.* Identifying the Hogue model

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?

The odds ratio estimates presented by Hogue roughly resemble Model 1c, the logistic regression model that includes the number of induced and spontaneous abortions, without adjustment for any covariates. Again, the absence of adjustment variables in their model is a clear limitation and ignorant to confounders.

9.* Critique of Trichopoulos and Hogue

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?

These estimates come from an outdated approach to modeling matched data. Although they are more reasonable than Model 4 (ordinary logistic regression with adjustment of matched sets), they do not use the matched-nature of their data to their advantage–which is fair as conditional logistic regression was not yet popularized.

11. Discussion

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.

- Strengths: Matching avoids confounding effects and matching at the individual level avoids residual confounding from matched variables. Matching by maternity department of the same hospital reduces potential influence from different healthcare factors. 1:2 matching improves statistical power. Both models including interaction effects grant greater flexibility in the estimated main effects.
- Limitations: Although the data followed a clever matching design, the estimates from Trichopoulos' and Hogue's models were naive to it, opening a door to biased conclusions. Additionally, the findings may be specific to hospitals included in the study if their sample is not representative. Finally, poorly or incorrectly defined levels of education may reduce the control over its confounding effect.

End of report. Code appendix begins on the next page.

Code Appendix

```
# clear environment
rm(list=ls())

# setup options
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, results='hide')
options(knitr.kable.NA = '-', digits = 2)
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "allcode")]
# load relevant packages
library(foreign)    # read data
library(survey)     # Wald test for a term in a regression model
library(survival)   # conditional logistic regression
library(car)        # test linear hypothesis

# load data
infert <- foreign::read.dta("../data/infert.dta")
ls(); names(infert); dim(infert)
str(infert)
head(infert)
# infert %>% group_by(stratum) %>% summarize(N=n()) %>% arrange(desc(N))

## handle missing data
anyNA(infert)  #no missing data

attach(infert)

############################
#### --- QUESTION 1 --- ####
############################

## 1a.
# Model secondary infertility using unadjusted ordinary logistic regression with dummy variables indica
induced.ctg <- as.factor(induced)
m1a <- glm(case ~ induced.ctg, family = "binomial")
summary(m1a)$coefficients

# Wald test for global significance of the number of prior induced abortions
survey::regTermTest(m1a, "induced.ctg", df = Inf)

## 1b.
# Model secondary infertility using unadjusted ordinary logistic regression with dummy variables indica
spont.ctg <- as.factor(spont)
m1b <- glm(case ~ spont.ctg, family = "binomial")
summary(m1b)$coefficients

# Wald test for global significance of the number of prior spontaneous abortions
regTermTest(m1b, "spont.ctg", df = Inf)

## 1c.
# Model secondary infertility using unadjusted ordinary logistic regression with dummy variables indica
m1c <- glm(case ~ induced.ctg + spont.ctg, family = "binomial")
```

```r
summary(m1c)$coefficients

# Wald test for global significance of the number of prior induced abortions
regTermTest(m1c, "induced.ctg", df = Inf)
# Wald test for global significance of the number of prior spontaneous abortions
regTermTest(m1c, "spont.ctg", df = Inf)

## 1d.
# Model secondary infertility using unadjusted ordinary logistic regression with dummy variables indica
m1d <- glm(case ~ g2 + g3 + g4 + g5 + g6 + g7 + g8, family = "binomial")
summary(m1d)$coefficients

# Wald test for global significance of eight combinations
car::linearHypothesis(m1d, c("g2 = 0", "g3=0", "g4=0", "g5=0", "g6=0", "g7=0",
                             "g8=0"), test = "Chisq")

# Likelihood ratio test for effect modification between number of induced abortions and number of sponta
anova(m1c, m1d, test = "Chisq")

############################
#### --- QUESTION 2 --- ####
############################

educ.ctg <- as.factor(educ)
parity.ctg <- as.factor(parity)

## 2a.
# Model secondary infertility using adjusted ordinary logistic regression with dummy variables indicati
m2a <- glm(case ~ induced.ctg + educ.ctg + age + parity.ctg, family="binomial")
rbind(summary(m1a)$coefficients,
      summary(m2a)$coefficients)

# Wald test for global significance of the number of prior induced abortions
regTermTest(m1a, "induced.ctg", df = Inf)
regTermTest(m2a, "induced.ctg", df = Inf)

## 2b.
# Model secondary infertility using adjusted ordinary logistic regression with dummy variables indicati
m2b <- glm(case ~ spont.ctg + educ.ctg + age + parity.ctg, family = "binomial")
rbind(summary(m1b)$coefficients,
      summary(m2b)$coefficients)

# Wald test for global significance of the number of prior spontaneous abortions
regTermTest(m1b, "spont.ctg", df = Inf)
regTermTest(m2b, "spont.ctg", df = Inf)

## 2c.
# Model secondary infertility using adjusted ordinary logistic regression with dummy variables indicati
m2c <- glm(case ~ induced.ctg + spont.ctg + educ.ctg + age + parity.ctg,
           family = "binomial")
rbind(summary(m1c)$coefficients,
      summary(m2c)$coefficients)
```

```r
# Wald test for global significance of the number of prior induced abortions
regTermTest(m1c, "induced.ctg", df = Inf)
regTermTest(m2c, "induced.ctg", df = Inf)
# Wald test for global significance of the number of prior spontaneous abortions
regTermTest(m1c, "spont.ctg", df = Inf)
regTermTest(m2c, "spont.ctg", df = Inf)
## 2d.
# Model secondary infertility using adjusted ordinary logistic regression with dummy variables indicati
m2d <- glm(case ~ g2 + g3 + g4 + g5 + g6 + g7 + g8 +
             educ.ctg + age + parity.ctg, family = "binomial")
rbind(summary(m1d)$coefficients,
      summary(m2d)$coefficients)

# Wald test for global significance of eight combinations
linearHypothesis(m2d, c("g2=0", "g3=0", "g4=0", "g5=0", "g6=0", "g7=0", "g8=0"),
                 test = "Chisq")

# Likelihood ratio test for effect modification between number of induced abortions and number of sponta
anova(m2c, m2d, test = "Chisq")
############################
#### --- QUESTION 3 --- ####
############################

## 3a.
# Model secondary infertility using unadjusted conditional logistic regression with dummy variables ind
m3a <- survival::clogit(case ~ induced.ctg + survival::strata(matchset))
summary(m2a)$coefficients
summary(m3a)$coefficients[1:2,]

# Wald test for global significance of the number of prior induced abortions
regTermTest(m2a, "induced.ctg", df = Inf)
regTermTest(m3a, "induced.ctg", df = Inf)

## 3b.
# Model secondary infertility using unadjusted conditional logistic regression with dummy variables ind
m3b <- clogit(case ~ spont.ctg + strata(matchset))
summary(m2b)$coefficients
summary(m3b)$coefficients

# Wald test for global significance of the number of prior spontaneous abortions
regTermTest(m2b, "spont.ctg", df = Inf)
regTermTest(m3b, "spont.ctg", df = Inf)

## 3c.
# Model secondary infertility using unadjusted conditional logistic regression with dummy variables ind
m3c <- clogit(case ~ induced.ctg + spont.ctg + strata(matchset))
summary(m2c)$coefficients
summary(m3c)$coefficients

# Wald test for global significance of the number of prior induced abortions
regTermTest(m2c, "induced.ctg", df = Inf)
regTermTest(m3c, "induced.ctg", df = Inf)
# Wald test for global significance of the number of prior spontaneous abortions
```

```r
regTermTest(m2c, "spont.ctg", df = Inf)
regTermTest(m3c, "spont.ctg", df = Inf)
## 3d.
# Model secondary infertility using unadjusted conditional logistic regression with dummy variables ind
m3d <- clogit(case ~ g2 + g3 + g4 + g5 + g6 + g7 + g8 + strata(matchset))
summary(m3d)$coefficients

# Wald test for global significance of eight combinations
linearHypothesis(m2d, c("g2=0", "g3=0", "g4=0", "g5=0", "g6=0", "g7=0", "g8=0"),
                 test = "Chisq")
linearHypothesis(m3d, c("g2=0", "g3=0", "g4=0", "g5=0", "g6=0", "g7=0", "g8=0"),
                 test = "Chisq")

# Likelihood ratio test for effect modification between number of induced abortions and number of sponta
anova(m2c, m2d, test = "Chisq")
anova(m3c, m3d, test = "Chisq")
## 4a.
# Model secondary infertility using adjusted ordinary logistic regression with dummy variables indicati
matchset.ctg <- as.factor(matchset)
m4a <- glm(case ~ induced.ctg + matchset.ctg, family = "binomial")
summary(m4a)$coefficients

# Wald test for global significance of the number of prior induced abortions
regTermTest(m4a, "induced.ctg", df = Inf)
## 4b.
# Model secondary infertility using adjusted ordinary logistic regression with dummy variables indicati
m4b <- glm(case ~ spont.ctg + matchset.ctg, family = "binomial")
summary(m4b)$coefficients

# Wald test for global significance of the number of prior spontaneous abortions
regTermTest(m4b, "spont.ctg", df = Inf)

## 4c.
# Model secondary infertility using adjusted ordinary logistic regression with dummy variables indicati
m4c <- glm(case ~ induced.ctg + spont.ctg + matchset.ctg, family = "binomial")
summary(m4c)$coefficients

# Wald tests for global significance of the number of prior abortions
regTermTest(m4c, "induced.ctg", df = Inf)
regTermTest(m4c, "spont.ctg", df = Inf)

## 4d.
# Model secondary infertility using adjusted ordinary logistic regression with dummy variables indicati
m4d <- glm(case ~ g2 + g3 + g4 + g5 + g6 + g7 + g8 + matchset.ctg,
           family = "binomial")
summary(m4d)$coefficients

# Wald test for global significance of eight combinations
linearHypothesis(m4d, c("g2=0", "g3=0", "g4=0", "g5=0", "g6=0", "g7=0", "g8=0"),
                 test = "Chisq")

# Likelihood ratio test for effect modification between number of induced abortions and number of sponta
anova(m4c, m4d, test = "Chisq")
```

```
############################
#### --- QUESTION 5 --- ####
############################

## 5a.
# Model secondary infertility using conditional logistic regression with pooled matched sets and with d
m5a <- clogit(case ~ induced.ctg + strata(stratum))
summary(m5a)$coefficients

# Wald test for global significance of the number of prior induced abortions
regTermTest(m5a, "induced.ctg", df=Inf)

## 5b.
# Model secondary infertility using conditional logistic regression with pooled matched sets and with d
m5b <- clogit(case ~ spont.ctg + strata(stratum))
summary(m5b)$coefficients

# Wald test for global significance of the number of prior spontaneous abortions
regTermTest(m5b, "spont.ctg", df=Inf)

## 5c.
# Model secondary infertility using conditional logistic regression with pooled matched sets and with d
m5c <- clogit( case ~ induced.ctg + spont.ctg + strata(stratum))
summary(m5c)$coefficients

# Wald tests for global significance of the number of prior abortions
regTermTest(m5c, "induced.ctg", df=Inf)
regTermTest(m5c, "spont.ctg", df=Inf)

## 5d.
# Model secondary infertility using conditional logistic regression with pooled matched sets and dummy
m5d <- clogit( case ~ g2 + g3 + g4 + g5 + g6 + g7 + g8 + strata(stratum))
summary(m5d)$coefficients

# Wald test for global significance of eight combinations
linearHypothesis(m5d, c("g2=0", "g3=0", "g4=0", "g5=0", "g6=0", "g7=0", "g8=0"),
                 test = "Chisq")

# Likelihood ratio test for effect modification between number of induced abortions and number of spont
anova(m5c, m5d, test = "Chisq")
```

End of document.