Exercise 5.1: Doubling the cases

In a population-based study of risk factors for hemolytic and non-haemolytic neonatal jaundice (Lee et al, Acta Paediatrika, 2016, DOI: 10.1111/apa.13470), the crude RR of haemolytic jaundice from 1992-2002 for singleton infants born preterm was 16.5, and the crude OR was 27.8.

The dataset **neonatal_jaundice.dta** contains a summary of these data, where the variable **num_neonates** provides the numbers of infants in each of the categories defined by the outcome of interest event (diagnosis of jaundice: **event**), the main exposure of interest (preterm delivery: **preterm**) and a few other variables (mother age 35 or older: **age35**, BMI category: **overobes** [0 =neither, 1= overweight, 2 =obese], multiparity: **multipar**, sex of neonate: **female**).

(i) Expand the data to one record per infant, tabulate exposure by outcome, and find the crude RR and crude OR for the association between preterm delivery and jaundice

(ii) Run a logistic regression to obtain the OR for preterm, adjusted for maternal age, BMI, parity and sex of infant. Verify that your adjusted OR is close to the value reported in the last column of the table at the end of this document (analysis of a case-control sample from the full data that is adjusted for maternal smoking).

(iii) Modify the dataset using the double-the-cases method. Tabulate exposure by outcome in the modified data, compare to the exposure-by-outcome table from (i) to verify that the frequencies are correct, i.e.

Original Data

	Case: Yes	Case: No
Exposed	a	b
Unexposed	c	d

n

Modified Data

	Case: Yes	Case: No
Exposed	a	$b+a$
Unexposed	c	$d+c$

$n*$

(iv) Run a logistic regression analysis of the modified data to obtain the adjusted RR for preterm and compare to the value reported in the first column of the table below from the full data.

(v) Select a 1:2 case-control sample from the cohort using simple random sampling of controls, and compare the estimates of the adjusted OR from logistic regression and adjusted RR from weighted logistic regression to the values obtained from the full cohort in (ii) and (iv) respectively.

(vi) Repeat the selection of the 1:2 case-control sample from part (v), but this time select controls that are frequency matched on maternal age and infant sex. Assign appropriate weights to cases and controls. Double the cases and verify that an unweighted logistic regression provides a biased estimate of the adjusted RR, but a weighted logistic regression (weighted by the inverse of the sampling fractions) provides an unbiased estimate.

(vii) To obtain standard errors for the estimates above, further calculations are required. These

are automatically provided if the data are analyzed in R using the "DoublingOfCases" package available from https://github.com/nyilin/DoublingOfCases. If you are using R, reanalyze the data now using this package with the **logit_db** function. For the Stata output, comment the invalid variance is too small or too large.

	Double the Cases		Logistic
	Cohort	**Case-control**	**Case-control**
	RR (95% C.I.)	RR (95% C.I.)	OR (95% C.I.)
Preterm	15.5 (15.1, 15.9)	15.4 (14.7, 16.1)	26.5 (24.6, 28.5)
BMI <18.5	1.03 (0.96, 1.12)	1.18 (1.06, 1.31)	1.22 (1.07, 1.4)
BMI 22.5-30	1.19 (1.16, 1.23)	1.19 (1.14, 1.25)	1.22 (1.15, 1.29)
BMI $\geq$ 30	1.44 (1.38, 1.50)	1.41 (1.33, 1.5)	1.47 (1.36, 1.59)
Multiparious	0.58 (0.56, 0.60)	0.59 (0.57, 0.61)	0.55 (0.53, 0.58)
Female infant	0.80 (0.78, 0.82)	0.8 (0.77, 0.83))	0.78 (0.74, 0.82)

TABLE 5.10: *Estimated adjusted RRs from doubling the cases for the analysis of association between preterm delivery and neonatal jaundice in a population-based cohort and in a case-control sample. Adjusted ORs from Standard logistic regression are included for comparison. In addition to adjustment for all factors shown, estimates are adjusted for maternal age and smoking status*