Exercise 2.1 Logistic regression

These exercises are based on the dataset set **"2.1_Framingham_exercises.dta",** which contains the individual records for the participants in the Framingham study who were summarised in the tables in Exercise 1.2. The data set includes the following variables:

sex (1=male, 2=female)
age (age, in years)
agecat (four age categories in 10-year intervals: 0=32-41, 1=42-51, 2=52-61, 3=62-71)
diabetes (treatment or glucose >=200mg/dL: 0=no diabetes, 1=diabetes)
anychd (0=any CHD during follow-up, 1=no CHD during follow-up)
prevchd (0=no prevalent coronary heart disease, CHD, 1=prevalent CHD)
prevhyp (0=no prevalent hypertension, 1=prevalent hypertension)

1. For individuals aged 42 and older (i.e. **agecat**=1,2,3):

   (i) For individuals aged 42 and older (i.e. **agecat**=1,2,3), reproduce the following counts for incident CHD (**anychd**) and diabetes (already used in Exercise 1.2)

	Age 42–51			Age 52–61			Age 62–71		
	Diabetes			Diabetes			Diabetes		
	Yes	No		Yes	No		Yes	No	
Yes CHD	1	32	33	4	74	78	7	69	76
No CHD	28	1553	1581	47	1221	1268	29	440	469
	29	1585	1614	51	1295	1346	36	509	545

   (ii) Run a logistic regression to obtain the crude OR and check it against your answer from Exercise 1.2.

   (iii) Estimate the adjusted OR and compare to your answer in Exercise 1.2.

   (iv) Add an interaction term to your model in (iii) and compare the results to your conclusion regarding heterogeneity in Exercise 1.2

2. The following exercise compares the 2-by-2 tables analysis with logistic regression

   (i) Reproduce the following tables describing the association between prevalent hypertension (**prevhyp**) and diabetes for the 3505 individuals aged 42 or older:

	Overall			Males			Females		
	Diabetes			Diabetes			Diabetes		
	Yes	No		Yes	No		Yes	No	
Yes HTN	67	1237	1304	25	520	545	42	717	759
No HTN	49	2152	2201	32	951	983	17	1201	1218
	116	3389	3505	57	1471	1528	59	1918	1977
	OR = 2.38			**OR = 1.43**			**OR = 4.14**		

   (ii) Verify that analysing the data in 3 tables above using a logistic model with interaction term produces the results shown on the next page (Note that the confidence intervals may differ slightly depending on the software and command/package that you use):

Example 2.5 (continued). *Framingham Heart Study: Prevalent Hypertension (X) and Diabetes (Y): Modification by Sex (Z)*

	Exposure	Odds Ratio	95% C.I.
Unadjusted	Hypertension	2.38	1.63 – 3.46
Interaction	Hypertension	1.43	0.838 – 2.44
Model	Sex (Female)	0.421	0.232 – 0.762
	Sex* Hypertension	2.90	1.33 – 6.33

(iii) How can the OR for females found in (i) be obtained from the output in (ii)?

3. (optional)
   - (i) Run a logistic regression analysis of diabetes in the cohort aged 40 and above and record the crude OR for the effect of prevalent CHD (variable name **prevchd**) on diabetes (variable name **diabetes**), also the ln(OR) and the intercept of the logistic model.
   - (ii) Select a case-control sample consisting of all cases and an equal number of controls (chosen randomly: use a "seed" so that you can reproduce your results if needed).
   - (iii) Re-run the logistic regression using the case-control sample, again recording the crude OR for the effect of prevalent CHD, also the ln(OR) and the intercept. Compare to your estimate in (i): the OR or ln(OR) should agree to within sampling variation, but the intercept will be very different.

HINTS

Ex 1: Stata immediate commands csi (cross-sectional data) and cci (case-control data)
Ex 3:
Ex 4: Test for trend
OpenEpi: Dose-Response command
Stata: tabodds command

Ex 5: Stata immediate commands csi (cross-sectional data) and cci (case-control data)

Ex 13: Stata tabodds command