The University of Hong Kong
School of Public Health

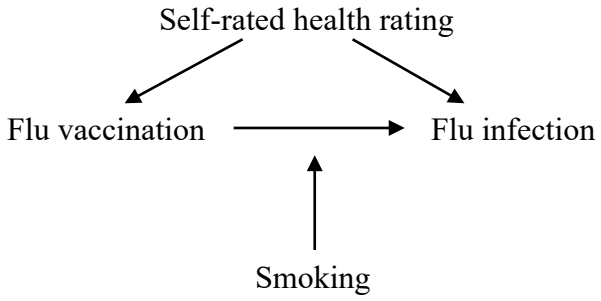
CMED 6040 Advanced Statistical Methods II
TUTORIAL 2

Influenza vaccine effectiveness: addressing missing data by multiple imputation

An observational study investigated the effect of influenza vaccination on influenza infection in the general population. To estimate the vaccine effect, potential confounding effects and effect modification will be considered. Data are available from 2000 adults aged 20 to 60y (saved in 'fluvaccine.csv'), with the following variables:

Variable	Data label	Remark
flu	Influenza infection during the flu peak season	0: No; 1: Yes
age	Age in years	
male	Male sex	0: Female; 1: Male
shealth	Self-rated health rating	0-10 (good health)
smoking	Current smoker	0: No; 1: Yes
bmi	Body mass index (kg/m^2)	
vac	Pre-seasonal influenza vaccination	0: No; 1: Yes
abT	Pre-seasonal antibody titer (but after vaccination)	by 2-fold serial dilution

(a) Read the dataset into R. Suppose that based on literature review and a preliminary analysis, the causal structure can be summarized by the following DAG:



Fit the above regression model and summarize the results in a table.

Variable	***Adjusted odds ratio***	***95% CI****
Flu vaccination		
Self-reported health rating		
Current smoker		
Flu vaccination × current smoker		

*CI: confidence interval

(b) Suppose the same dataset have some missing data ('fluvaccine_m1.csv'). Study the dataset especially on the missingness of the data.

(c) It was known that missingness was completely random (MCAR). Carry out a complete case analysis and compare the results with that in (a).

Variable	*Adjusted odds ratio*	*95% CI**
Flu vaccination		
Self-reported health rating		
Current smoker		
Flu vaccination × current smoker		

*CI: confidence interval

(d) To deal with missing data, multiple imputation method will be used. Create the imputed datasets (say, 50 imputations), based on the variables in the final analysis model. Assess if the variables were imputed sensibly by examining the imputed variables and comparing them with the original dataset.

(e) Perform the regression analysis using multiple imputation method. Compare your results with that in (c).

Variable	*Adjusted odds ratio*	*95% CI**
Flu vaccination		
Self-reported health rating		
Current smoker		
Flu vaccination × current smoker		

*CI: confidence interval

(f) Suppose the same dataset have data not missing completely at random (not MCAR, in 'fluvaccine_m2.csv'). Study the dataset especially on the missingness of the data.

(g) Carry out a complete case analysis and comparing the results with that in (a).

Variable	*Adjusted odds ratio*	*95% CI**
Flu vaccination		
Self-reported health rating		
Current smoker		
Flu vaccination × current smoker		

*CI: confidence interval

(h) Perform the regression analysis using multiple imputation method. Compare your results with those in (c) and (g).

Variable	*Adjusted odds ratio*	*95% CI**
Flu vaccination		
Self-reported health rating		
Current smoker		
Flu vaccination × current smoker		

*CI: confidence interval

(i) Suppose we make use of all available variables to create the imputed datasets. Perform the regression analysis and compare your results with those in (c), (g) and (h).

Variable	*Adjusted odds ratio*	*95% CI**
Flu vaccination		
Self-reported health rating		
Current smoker		
Flu vaccination × current smoker		

*CI: confidence interval

(j) Suppose the same dataset have more missing data (up to 30% missing for key variables, 'fluvaccine_m3.csv'). Study the dataset especially on the missingness of the data.

(k) Perform the regression analysis again using multiple imputation method. Compare your results with those in (c) and (h).

Variable	*Adjusted odds ratio*	*95% CI**
Flu vaccination		
Self-reported health rating		
Current smoker		
Flu vaccination × current smoker		

*CI: confidence interval