Interaction and confounding

April 8, 2014 Debdeep Pati

Modeling Binary outcome

Test of hypothesis

1. Is the effect observed statistically significant or attributable to chance?
2. Three types of hypothesis: a) tests of goodness of fit of the overall model. b) tests of effect of any one risk factor contained within the model. c) tests of the linear effect of ordered categorical risk factors.
3. Deviance is calculated from the likelihood, which is a measure of how likely a particular model is, given the observed data.
4. A measure of the difference between the postulated model and the model that, by definition, is a perfect fit to the data (called full or saturated model).
5. Deviance is given by
$$D = -2\{\log \hat{L} - \log \hat{L}_F\}$$
6. The deviance of the model can be used to test for goodness of fit of the model to the data. The model deviance is compared to chi-square with the model deviance df. The df for a model deviance is calculated as “df = number of data items - number of independent parameters in the fitted model”.
7. Number of independent parameters is 1 for the intercept term, 1 for quantitative variable and $l - 1$ for a categorical variable with l levels.
8. In the case of lack of fit, Further explanatory variables may be needed. We may have inadequately modeled the effect of the current variables. Transformations might be needed, important interactions might be missing, Outliers may be in the data. Assumption of binomial variation may be incorrect. It is much more meaningful to test for specific effects.

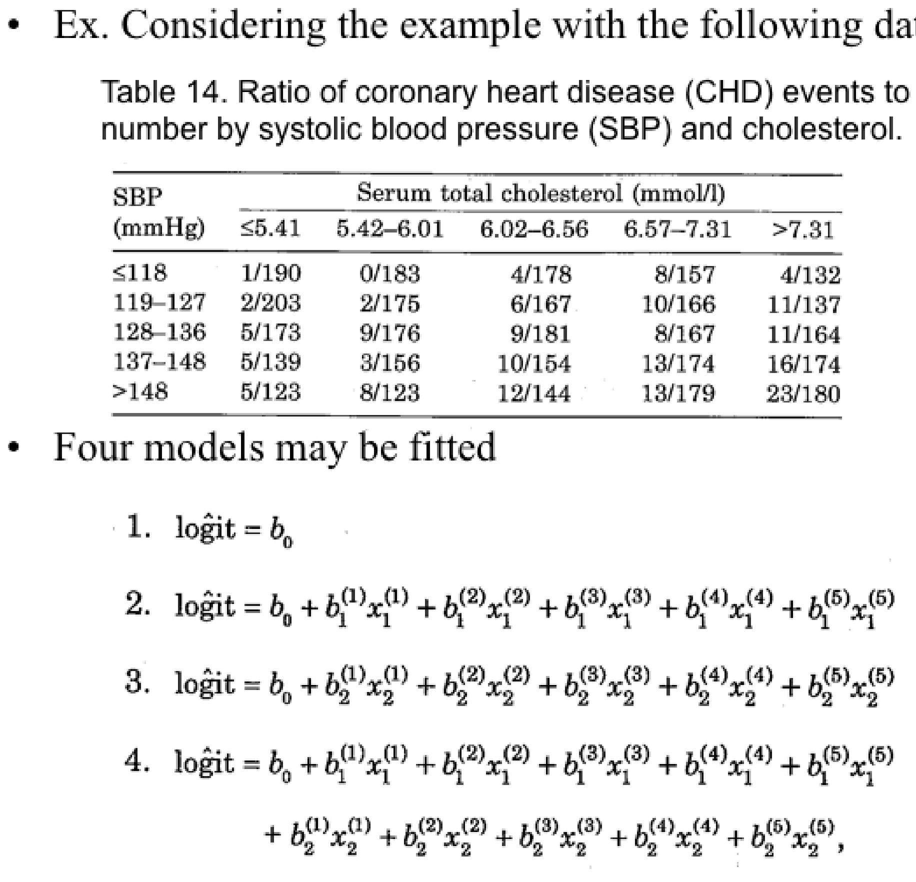
Effect of a Risk factor

Model nesting: Model A is said to be nested within model B if model B contains all the variables of model A plus at least one other. Constant is thought of as a variable.

Table 1: default

Model A	Model B
constant	constant + social class
constant + SBP	constant + SBP + cholesterol
constant + age + cholesterol + BMO + smoking	constant + age + cholesterol+ BMO + SBP + smoking+ activity in leisure

When model A is nested within model B, we can test the hypothesis that the extra terms in B have no effect by calculating the difference between the deviance of models A and B, denoted ΔD.



- Ex. Considering the example with the following data

Table 14. Ratio of coronary heart disease (CHD) events to total number by systolic blood pressure (SBP) and cholesterol.

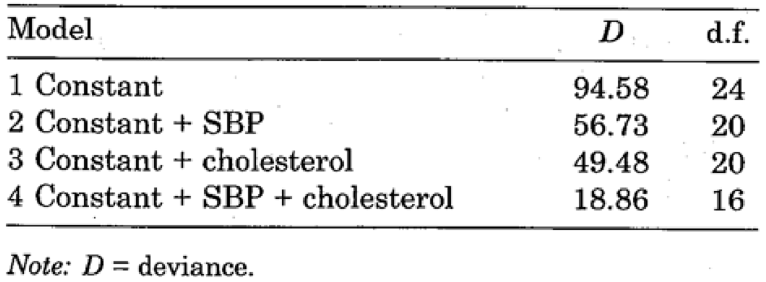
SBP	Serum total cholesterol (mmol/l)				
(mmHg)	≤5.41	5.42–6.01	6.02–6.56	6.57–7.31	>7.31
≤118	1/190	0/183	4/178	8/157	4/132
119–127	2/203	2/175	6/167	10/166	11/137
128–136	5/173	9/176	9/181	8/167	11/164
137–148	5/139	3/156	10/154	13/174	16/174
>148	5/123	8/123	12/144	13/179	23/180

- Four models may be fitted

1. $\hat{\text{logit}} = b_0$
2. $\hat{\text{logit}} = b_0 + b_1^{(1)}x_1^{(1)} + b_1^{(2)}x_1^{(2)} + b_1^{(3)}x_1^{(3)} + b_1^{(4)}x_1^{(4)} + b_1^{(5)}x_1^{(5)}$
3. $\hat{\text{logit}} = b_0 + b_2^{(1)}x_2^{(1)} + b_2^{(2)}x_2^{(2)} + b_2^{(3)}x_2^{(3)} + b_2^{(4)}x_2^{(4)} + b_2^{(5)}x_2^{(5)}$
4. $\hat{\text{logit}} = b_0 + b_1^{(1)}x_1^{(1)} + b_1^{(2)}x_1^{(2)} + b_1^{(3)}x_1^{(3)} + b_1^{(4)}x_1^{(4)} + b_1^{(5)}x_1^{(5)}$

$$+ b_2^{(1)}x_2^{(1)} + b_2^{(2)}x_2^{(2)} + b_2^{(3)}x_2^{(3)} + b_2^{(4)}x_2^{(4)} + b_2^{(5)}x_2^{(5)},$$

- Analysis of deviance table



Model	D	d.f.
1 Constant	94.58	24
2 Constant + SBP	56.73	20
3 Constant + cholesterol	49.48	20
4 Constant + SBP + cholesterol	18.86	16

Note: D = deviance.

- Compare models 1 and 2 to assess the significance of SBP.
- Models 1 and 3 for cholesterol
- Models 1 and 4 for SBP and cholesterol together
- Models 3 and 4 for SBP over and above cholesterol
- Models 2 and 4 for cholesterol over and above SBP.

Confounding and Interaction

We may be concerned with only two variables, such as a risk factor and disease status. If the third factor can explain (at least partially) the relationship of the two variables, then *c*onfounding is present. e.g. Relationship between the number of children and probability of breast cancer may be explained by the ages of the mothers. If the third factor modifies the relationship between risk factor and the disease, then *interaction* is present. e.g. Relationship between salt consumption and stroke is quite different for men and women. Then gender interacts with salt consumption in determining the risk of a stroke.

Definition of a confounder

Confounder (a confounding variable) is an an extraneous factor that wholly or partially accounts for the observed effect of the risk factor on disease status. There are two scenarios for effects.

1. an apparent relationship: the confounder is causing the relationship to appear.
2. an apparent lack of relationship: the confounder is masking a true relationship.

Assessing confounding

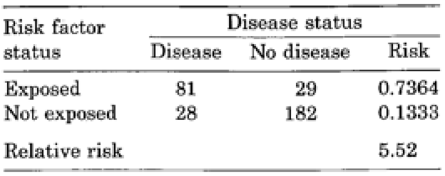
Confounding

1. Adjustment for confounding variables is achieved through logistic modeling by fitting the confounder with and without the risk factor.
2. Comparison of odds ratios from the models with the risk factor alone and with the confounder added indicates the effect of the confounder.

Interaction

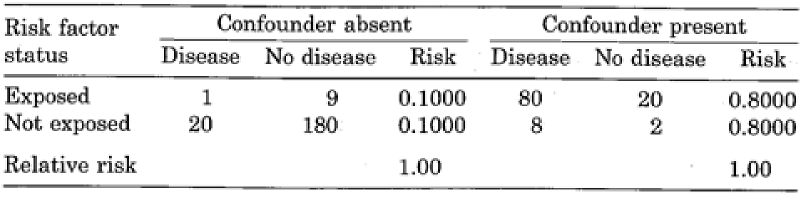
1. Interaction is dealt with by introducing one or more terms into the logistic regression model.
2. Between two categorical variables, Between a quantitative and a categorical variable, Between two quantitative variables.
3. Whenever an interaction turn out to be significant, the main effect of the constituent terms are likely to be misleading.

Table 1. Risk factor status by disease status



Risk factor	Disease status		
status	Disease	No disease	Risk
Exposed	81	29	0.7364
Not exposed	28	182	0.1333
Relative risk			5.52

Table 2. Risk factor status by disease status by confounder (C) status



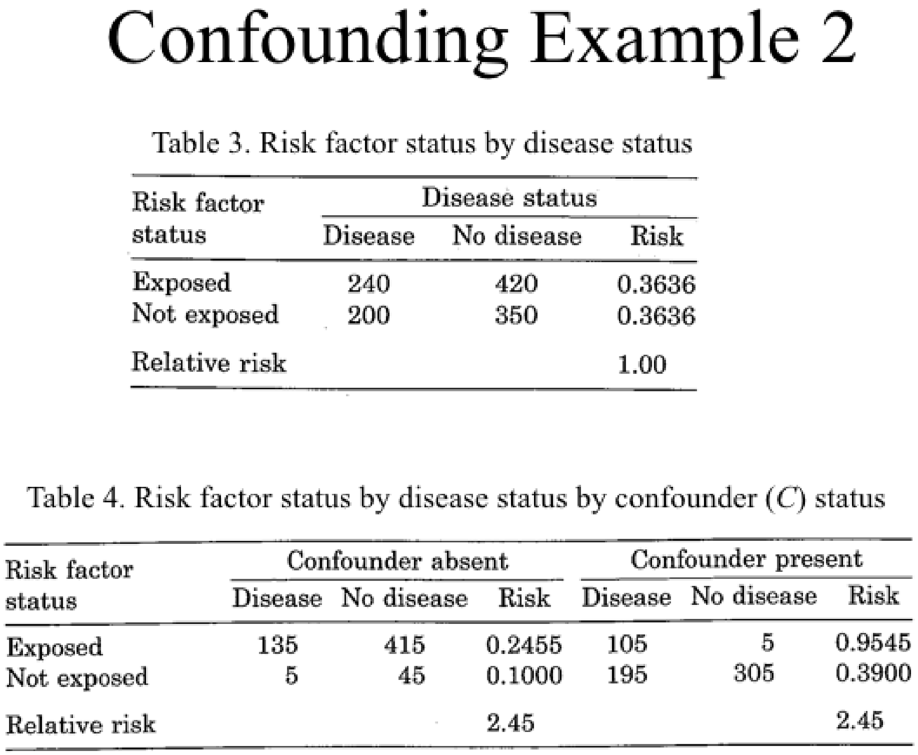
Risk factor	Confounder absent			Confounder present		
status	Disease	No disease	Risk	Disease	No disease	Risk
Exposed	1	9	0.1000	80	20	0.8000
Not exposed	20	180	0.1000	8	2	0.8000
Relative risk			1.00			1.00

Reasons for confounding

1. presence/absence of the confounder and the risk factor tend to go together.

2. C is itself, a risk factor for the disease. $RR = \frac{(80+8)/(80+8+20+2)}{(1+20)/(1+20+9+180)} = 8$.

Example 2



Confounding Example 2

Table 3. Risk factor status by disease status

Risk factor status	Disease	No disease	Risk
Exposed	240	420	0.3636
Not exposed	200	350	0.3636
Relative risk			1.00

Table 4. Risk factor status by disease status by confounder (C) status

Risk factor status	Confounder absent			Confounder present		
	Disease	No disease	Risk	Disease	No disease	Risk
Exposed	135	415	0.2455	105	5	0.9545
Not exposed	5	45	0.1000	195	305	0.3900
Relative risk			2.45			2.45

1. The presence of the confounder tends to go with the absence of the risk factor whilst the absence of the confounder tends to go with the presence of the risk factor.

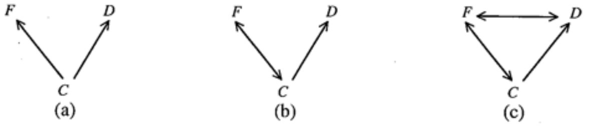
2. C is, itself, a risk factor with relative risk $RR = \frac{(105+195)/(105+195+5+305)}{(135+5)/(135+5+415+45)} = 2.11$.

1 Identification of confounders

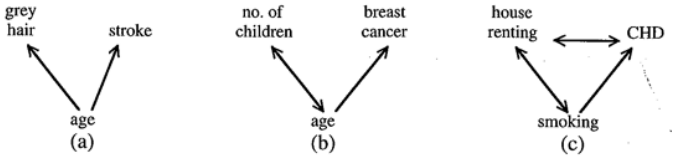
1. D, disease, F, risk factor, and C, the third variable

2. If C is a confounder, it must i) either be related to the disease, but not a consequence of the disease. ii) or be related to the risk factor, but not a consequence of the risk factor.

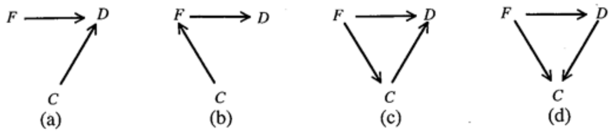
3. Path diagrams:
   i) arrows show relationships that exist regardless of all other relationships.
   ii)double-sided arrows are used to denote noncausal relationships.
   iii) single-sided arrows show the direction of causality.



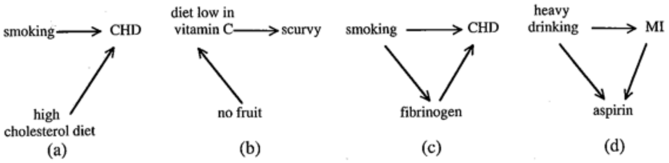
Some situations in which *C* is a confounder for the *F*–*D* relationship.



Some examples that may fit the situations

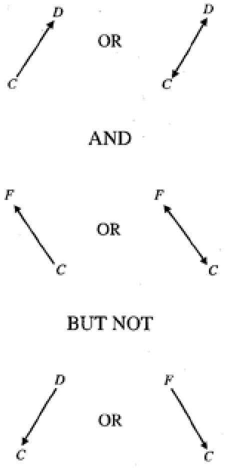


Some situations in which C is not a confounder for the F–D relationship.



Some examples that may fit the situations

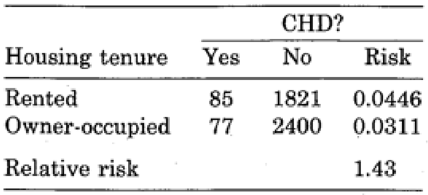
Strategy for Selection



Conditions for C to be a confounder for F-D relationship

Confounding Example 3

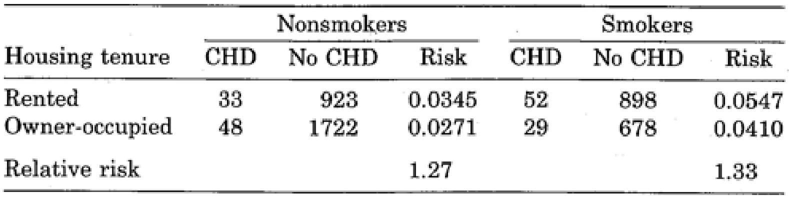
Table 5. Housing tenure by CHD outcome after 6 years; SHHS men



	CHD?		
Housing tenure	Yes	No	Risk
Rented	85	1821	0.0446
Owner-occupied	77	2400	0.0311
Relative risk			1.43

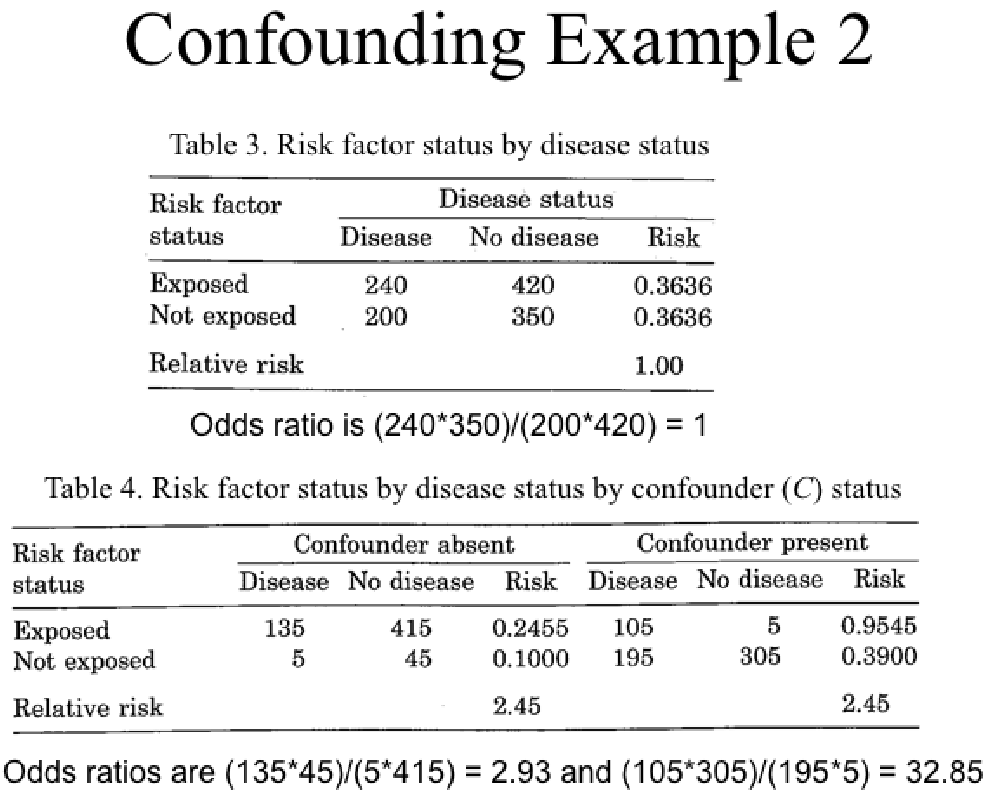
6 years' follow-up of men in the Scottish Heart Health Study (SHHS). These data are for those with no symptoms of coronary heart disease (CHD) at the beginning of the study. The variable 'housing tenure' records whether they rent or own their accommodation.

Table 6. Housing tenure by CHD outcome by cigarette smoking after 6 years.



	Nonsmokers			Smokers		
Housing tenure	CHD	No CHD	Risk	CHD	No CHD	Risk
Rented	33	923	0.0345	52	898	0.0547
Owner-occupied	48	1722	0.0271	29	678	0.0410
Relative risk			1.27			1.33

Assess confounding by estimating the effect of the risk factor with and without allowing for confounding In the earlier example, the relative risk of renting is 1.43 unadjusted, and around 1.30 after adjustment for smoking. The effect of confounding can be estimated as Ec/E, where E is the unadjusted and Ec is the adjusted, estimate. 1.30/1.43 = .91, adjustment has reduced the relative risk by 9%. This approach depends on the risk measure used. When odds ratio is used the results can be quite different with when relative risk is used. For rare disease, odds ratio give similar results (Miettenen, OS and Cook, EF (1981) Confounding: essence and detection. Am J Epidemiol. 114, 593-603)



Confounding Example 2

Table 3. Risk factor status by disease status

Risk factor status	Disease status		
	Disease	No disease	Risk
Exposed	240	420	0.3636
Not exposed	200	350	0.3636
Relative risk			1.00

Odds ratio is (240*350)/(200*420) = 1

Table 4. Risk factor status by disease status by confounder (*C*) status

Risk factor status	Confounder absent			Confounder present		
	Disease	No disease	Risk	Disease	No disease	Risk
Exposed	135	415	0.2455	105	5	0.9545
Not exposed	5	45	0.1000	195	305	0.3900
Relative risk			2.45			2.45

Odds ratios are (135*45)/(5*415) = 2.93 and (105*305)/(195*5) = 32.85