Biostats VHM 811 Course
Fall 2008, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment 2

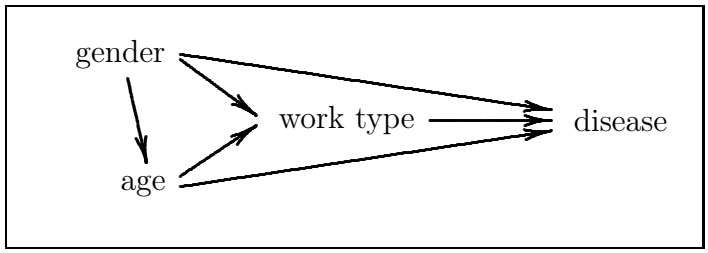
The data are presented in Healy (1980): *GLIM: An Introduction*, and are described to refer to the Finistére region of France. The disease is monoclonal gammapathy (or gammopathy), and the work type is farming (`work=0`) and non-farming (`work=1`) groups. but their exact origin is unknown. The dataset is similar to and most likely linked to the data published in Saleun *et al.* (1982), Monoclonal gammopathies in the adult population of Finistère, France, *J. Clin. Path.* **35**, 63–68.

The solution is more detailed than expected for a 100% mark.

1. Study design and causal diagram

The description of the study is very limited and does not by itself clearly determine the study type. Without any mention of a follow-up period it seems unlikely to have been a cohort study. The small number of cases (diseased persons) compared to the number of non-cases does not suggest a case-control study, as these two groups are typically selected to be of similar (if not the same) size. A survey or cross-sectional study seems the most likely study type. Indeed, the Saleun *et al.* study was reported as a cross-section: "In an attempt to assess the frequency of monoclonal gammopathy in a representative cross-section of the population we examined ...".

The causal relationships between the three variables can be hypothesized based on the information that "work type" is the exposure of interest. Age and gender must be considered as antecedants to work type, and all three variables should have a potential effect on disease. Among age and gender the most natural causal direction is perhaps from gender to age; the age distribution among men and women is different, because women tend to live longer. These considerations lead to the causal diagram shown below.



2. Crude measures of association with disease

For a cross-sectional study, both relative risk (RR) and odds-ratio (OR) are appropriate measures of association. It is also valid to use the risk difference, but as some of the following calculations are restricted to RR and OR these will be the focus here. Disease is rare, thus RR and OR are expected to be close. For the binary predictors work type and gender, calculation of the RR and OR with associated confidence intervals and Pearson chi-square significance tests is straightforward; the values are shown in the table below.

	Relative risk		Odds-ratio		Pearson test	
Predictor	RR	95% CI	OR	95% CI	X^2	P-value
work type	0.336	(.214,.528)	0.334	(.213,.523)	24.85	<.00005
gender	0.637	(.511,.794)	0.634	(.508,.792)	16.40	.0001

Based on the crude associations, both work type and gender are protective factors with significant associations with disease. Females are less likely to be diseased, as are persons with work type 1.

For the categorical predictor age group, the Pearson chi-square test still applies but there is no longer a single RR or OR. The two options are (i) to report RR or OR for each category relative to a "baseline" category, or to dichotomize the variable at one cut-point to obtain values comparing persons younger and older than the cutpoint; this calculation should be preferably be carried out at several cut-points to avoid losing information. For simplicity, the lowest age group has been chosen as the baseline category, although it is generally best to choose a category with a fairly large number of observations as the baseline category. See the table for results.

	Age group				
Parameter	0-39	40-49	50-59	60-69	70+
relative risk vs baseline	1	5.04	10.3	23.2	42.7
95% CI	n/a	(1.74,14.6)	(3.73,28.7)	(8.53,63.0)	(15.8,115.0)
odds-ratio vs baseline	1	5.06	10.4	23.5	44.0
95% CI	n/a	(1.82,14.0)	(3.90,27.8)	(8.98,61.7)	(16.9,114.4)
relative risk at cutpoint	n/a	19.3	7.80	5.71	4.32
95% CI	n/a	(7.20,51.7)	(5.23,11.6)	(4.41,7.39)	(3.49,5.35)
odds-ratio at cutpoint	n/a	19.5	7.92	5.82	4.42
95% CI	n/a	(7.55,50.5)	(5.32,11.8)	(4.49,7.54)	(3.56,5.50)
test	$X^2 = 285.2,\ \text{df} = 4,\ P < .0005$				

The association between age and disease is very strong (and significant). The measures against the lowest age group show increasing risks with increasing age. The measures obtained by dichotomizing age show dicreasing values at increasing cut-points. This reflects the same pattern in the data: that the risk in the lowest age group is very small relative to higher age categories (so when this category is pooled with higher categories, the measures drop in value).

3. Effect of gender on the relation between work type and disease

The causal diagram shows that gender size could be a confounder for work type. As a first step we carry out a Mantel-Haenszel (M-H) analysis. The homogeneity test is clearly non-significant: $X^2 = 0.20$ and $P = 0.66$ for RR ($X^2 = 0.21$ and $P = 0.65$ for OR). Therefore, there is no interaction efect. The Mantel-Haenszel estimate differs only moderately from the crude estimate: $0.301 \sim 10.5\%$ relative change for RR ($0.298 \sim 10.7\%$ for OR). With these findings, there is no (pressing) need to continue the analysis for confounding. There is no (substantial) confounding effect of gender on the relation between work type and disease.

4. Effect of age on the relation between work type and disease

Again, the causal diagram justifies an analysis for confounding by age, and we start by the M-H analysis. The homogeneity test is clearly non-significant: $X^2 = 1.62,\ P = 0.81$ for both RR and OR. Therefore, there is no interaction efect. However, the M-H estimates deviate strongly from the crude estimate: 0.993 ($\approx 200\%$ relative change) for both RR and OR. Also, the M-H estimate is not significantly different from zero (as assessed by its confidence interval or the chi-square test), in contrast to the crude estimates. There are two further conditions to check for whether age is indeed a confounder:

1. *Association with outcome.* The crude associations were computed in Question 2, but we need to consider an association that is not indirectly caused by the exposure. According the VER guidelines (p. 239), this means an assessment in the exposure-negative group. This approach should be valid also for a cross-sectional design. Without any information about the meaning of the work type variable, the category corresponding to lack of exposure can be determined by looking at the crude and adjusted associations. By the crude associations, `work=1` has the lowest risk, whereas by the M-H estimate the risk is about equal. If anything, exposure-negative should therefore refer to the `work=1` group. One may also assess the association in both groups of work type. Chi-square and Fisher's exact tests confirm the presence of an association in both groups, in the same direction as the crude association. For the `work=1` groups, the values are $X^2 = 27.0$, df $= 4$, $P < 0.0005$ by both the χ^2-approximation and Fisher's exact test.

2. *Association with exposure.* The appropriate approach is not explained in VER for a cross-sectional study. The general rule is that the association between confounder and exposure should not derive from the association between exposure and disease. In a cohort study, the outcome (disease) is determined after sampling the exposed and non-exposed groups, therefore the crude association between confounder and exposure suffices. In a cross-sectional study the crude association may reflect an association between exposure and disease, and one should instead look at the association in the non-diseased group (same approach as a case-control study). For these data, the diseased group is very small, and therefore the two approaches will be numerically very similar. The association is very strong (and significant): $X^2 \approx 3600$, $P < 0.0005$. The distribution between work types is fairly balanced for age$<$40, and gets more and more skewed towards work type 0 for higher age groups.

We conclude that age meets all the conditions to be a confounder for the effect of work type on disease.

5. Combined effect of age and gender on the relation between work type and disease

In order to study the combined effect of age and gender, we need to look at the combined age and gender classifications into $2 \cdot 5 = 10$ categories. A M-H analysis stratified on the combined age and gender classification gives an effect estimate of 0.875 for RR (0.873 for OR), clearly non-significant homogeneity tests, and confidence intervals for the stratified estimate that extend far on both sides of 1. These results are not substantially different from those obtained when stratifying on age alone (Question 4). The M-H estimate has dropped slightly (from 0.99 to 0.87), but there is no indication that this drop signals anything but random variation. Gender itself had no (or minimal) confounding effect on the relation. Therefore, if stratifying on age and gender in combination is no (or at most, little) different than stratifying on age alone, gender cannot be said to modify the effect of age on the relation between work type and disease.

The M-H estimate for work type after stratification on age was 0.993 (both RR and OR) and absolutely non-significant. Thus, age is an antecedent variable with complete confounding for work type (VER Section 13.11.4). Gender did not exert any (substantial) confounding on work type, neither on its own or in combination with age. It is of no real interest to characterize the effect of gender further in terms of the epidemiological scenarios in Section 13.11 because its effect is clearly secondary to the role of age.

6. Relationship of age and gender with disease

The previous analysis determined that there was no effect of work type on disease. Work type will therefore be ignored for all calculations here, and the resulting causal diagram is obtained by

removing work type from the diagram in Question 1. According to the diagram, gender might act as a confounder for the effect of age, but the converse is not possible. We will examine the associations of the diagram in turn.

1. *Association of age with outcome.* The five categories of age poses a problem here because our emphasis in the course has been on dichotomous exposures. We will dichotomize age at each of the four cutpoints from Question 2, and carry out a Mantel-Haenszel analysis stratified on gender at each cutpoint.

		Cutpoint for age				
Parameter	Statistic	0-39	40-49	50-59	60-69	70+
relative risk	M-H estimate	n/a	19.9	8.01	5.91	4.42
	crude estimate	n/a	19.3	7.80	5.71	4.32
	relative change	n/a	3%	3%	4%	2%
	P for homog. test	n/a	0.70	0.62	0.64	0.40
odds-ratio	M-H estimate	n/a	20.2	8.14	6.03	4.53
	crude estimate	n/a	19.5	7.92	5.82	4.42
	relative change	n/a	3%	3%	4%	2%
	P for homog. test	n/a	0.70	0.64	0.62	0.37

   The M-H and crude estimates differ only little; there is no confounding effect of gender on the relation between age and disease. All measures are significantly different from 1, indicating the (obvious) presence of an age effect.

2. *Association between gender and disease.* Following the same procedure as in Question 4 we should assess the association between gender and age among the non-exposed subjects. This is not really meaningful here because age has multiple categories and seems to exert a gradual effect. One could perhaps use the lowest age group as non-exposed. We will instead carry out a stratified analysis on age, thereby assessing the association for all age groups. There is no evidence of heterogeneity between age groups with a P-value for the chi-square homogeneity test above 0.9 for both RR and OR. The crude and stratified estimates are close at 0.64 and 0.57, respectively, for RR (and almost identical for OR). The M-H estimates are strongly significant (different from unity). The association between gender and disease seems to be evident.

3. *Association between gender and age.* As in Question 4, we assess the association in the non-diseased group. The Pearson chi-square test value is 154.6 with df=4 and very clearly significant. At inspection of the crosstable for gender and age it is obvious that females are underrepresented in lower age groups (46% for age$<$40) but overrepresented in higher age groups (55% for age$\geq$ 70). The association between gender and age is clear.

In summary, all three associations in the causal diagram for gender, age and disease exist, and at substantial strength. Nevertheless, gender does not act as a confounder for the age effect. The epidemiological interpretation is that gender is an antecedent variable (for age) with partial confounding (Section 13.11.5), where indeed the partial confounding is minimal. Conversely, age is an intervening variable for gender. The effects of age and gender may be quantified by crude associations (given in Question 2).