Biostatistics (MATH11230)

Vanda Inácio

University of Edinburgh

Semester 1, 2021/2022

Assessment of confounding using logistic regression models

General context

- In the following methodological exposition, we follow Chapter 15 of Jewell (2003) very closely.

- For simplicity, we focus on two risk factors, say X_1 and X_2, and assume that there is no (multiplicative) interaction between X_1 and X_2.

- The ideas presented throughout generalise directly to the situation where there are more than two risk factors being analysed.

- Let us consider the following two logistic regression models:

$$\log\left(\frac{p_{x_1}}{1-p_{x_1}}\right) = \beta_0 + \beta_1 x_1, \tag{1}$$

$$\log\left(\frac{p_{x_1,x_2}}{1-p_{x_1,x_2}}\right) = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \beta_2 x_2. \tag{2}$$

- Here $p_{x_1} = \Pr(D \mid X_1 = x_1)$ and $p_{x_1,x_2} = \Pr(D \mid X_1 = x_1, X_2 = x_2)$.

Assessment of confounding using logistic regression models

General context

- ↪ As the notation suggests, there is a distinction between the two slope coefficients, β_1 and $\tilde{\beta}_1$, associated with the variable X_1 in these models.

- ↪ Because X_2 is not involved in the model (1), the coefficient β_1 is interpreted as the log odds ratio associated with a unit increase in the scale of X_1, ignoring any potential confounding role of X_2 (or any other variable, for that matter).

- ↪ In turn, $\tilde{\beta}_1$ in model (2) gives the log odds ratio associated with a unit increase in X_1, *holding the variable X_2 fixed*.

- ↪ The log odds ratio given by $\tilde{\beta}_1$ in model (2), measures the effect of a unit increase in X_1, in any population group that shares a common value of X_2.

- ↪ Hence, this measure of association of X_1 and D directly accounts for the possible confounding effect of X_2.

Assessment of confounding using logistic regression models

General context

- ↪ Note that for simple categorical exposures, the maximum likelihood estimate of $\tilde{\beta}_1$ is an alternative estimator to the Mantel-Haenszel approach for estimating a common (log) odds ratio, accounting for the presence of a possible confounding variable.
- ↪ Maximum likelihood is a slightly different technique and so will usually yield similar, but not identical, results to the Mantel-Haenszel approach.
- ↪ Further, for arbitrary exposure variables, comparison of the estimates of the two parameters, β_1 and $\tilde{\beta}_1$, allows us to assess the confounding of the X_1–D relationship induced by X_2, assuming that models (1) and (2) adequately describe the exposure effects.

Assessment of confounding using logistic regression models

General context

↪ If the maximum likelihood estimates $\widehat{\beta}_1$ and $\widehat{\tilde{\beta}}_1$, are very similar, then there is little confounding of the X_1–D relationship by X_2.

↪ If the estimates are markedly different, then the data indicate that X_2 has an important confounding effect, and we should therefore account for X_2 in our analysis.

↪ As a rule of thumb, and as in the Mantel-Haenszel approach, if the ratio (in terms of the odds ratio)

$$\frac{e^{\widehat{\beta}_1} - e^{\widehat{\tilde{\beta}}_1}}{e^{\widehat{\beta}_1}},$$

is larger than 10% in size, then most epidemiologists would agree that there is evidence of confounding.

Assessment of confounding using logistic regression models

Example

- ↪ As an example, we use the Western Collaborative Group Study data (that arises from a population-based design).

- ↪ We will study, this time, the effect of smoking on the risk of coronary heart disease (CHD).

- ↪ The variable `Ncigs0` in the dataset has information on the number of cigarettes smoked per day by each individual. To keep matters simple, we create a binary variable, that takes the value 0 if the individual does not smoke any cigarettes per day and 1 otherwise.

- ↪ After estimating the logistic regression model with binary smoking status as the only exposure, we obtain that the estimate of the corresponding regression coefficient is 0.6299.

- ↪ This means that smokers have $e^{0.6299} = 1.877$ times the odds of CHD than non-smokers. The corresponding 95% CI is $(1.445, 2.440)$.

Assessment of confounding using logistic regression models

Example

- ↪ Now, let us averiguate if the estimated association between smoking status and CHD was confounded by age.
- ↪ From a conceptual perspective, age satisfies the three criteria for being a potential confounder variable:
  - ↪ Age is likely a predictor of CHD. As age increases, the likelihood of CHD probably increases.
  - ↪ Age could also be positively or negatively associated with smoking, depending on the population.
  - ↪ It does not make sense for age to be in the causal pathway between smoking and CHD.

Assessment of confounding using logistic regression models

Example

↪ Let us check, based on the data, the bivariate associations between age and smoking status and age and CHD.

↪ We can use t-tests or look at density plots (or boxplots) of the distributions.

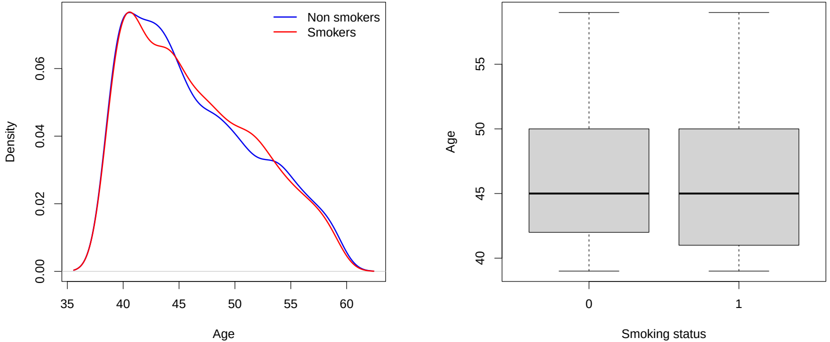
↪ Let us start with the association between age and smoking. We can create two groups of age: one for smokers and another one for non-smokers.

↪ A t-test comparing the two age groups leads to not reject the null hypothesis that the means of the two groups are the same. The mean age is similar for smokers and non-smokers.

Assessment of confounding using logistic regression models

Example

↪ The densities/boxplots in the two groups are basically indistinguishable.



Assessment of confounding using logistic regression models

Example

- ↪ We have just seen that for this specific dataset age does not seem to be associated with smoking status.

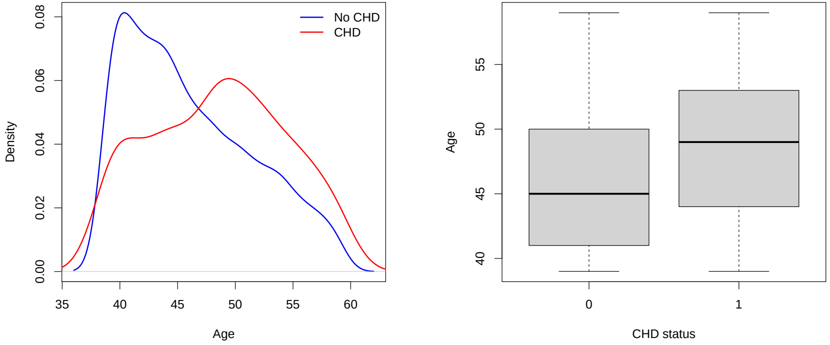
- ↪ Neverthless, let us proceed and check whether age is associated with the CHD outcome.

- ↪ By creating two groups of age, one for those with CHD and another one for those without CHD, the corresponding t-test leads us to reject the null hypothesis of equal means in the two groups.

- ↪ There is a significant difference in mean age by CHD status. Those with CHD tend to be older than those without CHD.

Assessment of confounding using logistic regression models

Example



Assessment of confounding using logistic regression models

Example

- ↪ To sum up: age was significantly associated with CHD but not with smoking.
- ↪ Let us look now at the change in the smoking coefficient when age is added to the model.
- ↪ Recall that in our simple logistic regression model the estimated regression coefficient associated with smoking status was 0.6299 and its estimated standard error was 0.1337.
- ↪ When age is added to the model, the estimated smoking coefficient is 0.63816 and the estimated standard error is 0.13472. The corresponding odds ratio is 1.893.
- ↪ There is not much of a change compared with the unadjusted (for age) odds ratio. Further, the aforementioned rule of thumb is only about 1% in size.
- ↪ So based on the lack of odds ratio change and on the bivariate associations, it does not look like age is a confounder.

Assessment of confounding using logistic regression models

Example

- ↪ But we should also check if including age makes our model better by increasing predictive power.
- ↪ This could provide some evidence to keep age in the model!
- ↪ We can test this by conducting a likelihood ratio test (the reduced model is nested within the full model) or by looking at the AIC/BIC.
- ↪ The AIC/BIC of the second model is lower than the AIC/BIC of the first one, suggesting that the second model is better.
- ↪ The null hypothesis of the likelihood ratio test is that the regression coefficient associated with age is zero. The p-value obtained is basically zero, and so there is strong evidence in favour of rejecting the null hypothesis.
- ↪ We would arrive at the same conclusion by doing a Wald test.

Assessment of confounding using logistic regression models

Example

- ↪ So, what do we do? Should we keep age in the model or not?
- ↪ We could argue either way!!
- ↪ We saw that age did not meet the criteria for confounding numerically and the smoking coefficient did not change much. Based on these reasons, we might not want to keep age in the model.
- ↪ But we saw that conceptually age could be a potential confounder, and we also see from the likelihood ratio test and AIC/BIC that it improved the model. Often, age is also adjusted for in epidemiological studies for face validity. These are reasons to keep it in the model.
- ↪ Personally, and for the reasons mentioned in the previous point, I would keep age in the model.