GLM Practical 6: Solutions

Part A: Kidney stones example

In session 5 and practical 5 we estimated the marginal and conditional treatment effects (of X on Y) using risk differences, risk ratios and odds ratios. See Tables 5.2 and 5.3.

1. In the notes we saw that the marginal risk difference is equal to a weighted average of the conditional risk differences (conditional on stone size Z). What are the weights? Show that the marginal risk difference is equal to a weighted average of the conditional risk differences.

   It was found that the conditional risk difference in the $Z = 0$ group is 0.064 and the conditional risk difference in the $Z = 1$ group is 0.043. Using standardization we also found that the marginal risk difference is 0.054. According to the results in the notes, the marginal risk difference (RD) is a weighted average of the conditional risk differences (RD_0 and RD_1):

   $$RD = RD_0 \times \Pr(Z = 0) + RD_1 \times \Pr(Z = 1)$$

   In this data set we have $\Pr(Z = 0) = 0.51$ and $\Pr(Z = 1) = 0.49$. The weighted average is $0.064 \times 0.51 + 0.043 \times 0.49 = 0.054$.

2. In the notes we saw that the marginal risk ratio is equal to a weighted average of the conditional risk ratios (Conditional on stone size Z). What are the weights? Show that the marginal risk ratio is equal to a weighted average of the conditional risk ratios.

   The marginal risk ratio is a weighted average of the conditional risk ratios given $Z = 0, 1$, with weights $w_0 = \Pr(Y = 1|do(X = 0), Z = 0)\Pr(Z = 0)/\Pr(Y = 1|do(X = 0))$ and $w_1 = \Pr(Y = 1|do(X = 0), Z = 1)\Pr(Z = 1)/\Pr(Y = 1|do(X = 0))$.

   The estimated conditional risk ratios in the $Z = 0$ and $Z = 1$ groups are $RR_0 = 1.933$ and $RR_1 = 1.158$ respectively. For the weights we can estimate $\Pr(Y = 1|do(X = 0), Z = 0)$ by assuming $\Pr(Y = 1|do(X = 0), Z = 0) = \Pr(Y = 1|X = 0, Z = 0)$, i.e. that Z is the only confounder of the association between X and Y. Under the same assumption, we can use standardization to estimate $\Pr(Y = 1|do(X = 0))$. The estimates of the weights are $\hat{w}_0 = 0.210$ and $\hat{w}_1 = 0.790$. The weighted average is therefore

   $$1.933 \times 0.210 + 1.158 \times 0.790 = 1.32,$$

   which is the marginal risk ratio estimate, as found in practical 5.

3. Can the marginal odds ratio be expressed as a weighted average of the conditional odds ratios? Explain your answer.

No, because odds ratios are non-collapsible. We can show this algebraically (you were not expected to do this for this question).

Consider a logistic regression model for Y given X and Z:

$$\Pr(Y|X,Z) = \frac{e^{\beta_0+\beta_X X+\beta_Z Z}}{1+e^{\beta_0+\beta_X X+\beta_Z Z}}$$

with X and Z independent, so there is no confounding. Supposing that Z is binary for simplicity, to find $\Pr(Y|X)$ we can appeal to the law of total expectation (standardization), which (applied here) says that

$$\begin{aligned}
\Pr(Y|X) &= \Pr(Y|X,Z=1)\Pr(Z=1) + \Pr(Y|X,Z=0)\Pr(Z=0) \\
&= \frac{e^{\beta_0+\beta_X X+\beta_Z}}{1+e^{\beta_0+\beta_X X+\beta_Z}} p_Z + \frac{e^{\beta_0+\beta_X X+\beta_Z}}{1+e^{\beta_0+\beta_X X}}(1-p_Z)
\end{aligned}$$

where $p_Z = \Pr(Z=1)$. The expression for $\Pr(Y|X)$ cannot be expressed in terms of a logistic model, i.e. we *cannot* write it in the form

$$\Pr(Y|X) = \frac{e^{\beta_0^*+\beta_X X}}{1+e^{\beta_0^*+\beta_X X}}$$

So, after averaging/marginalizing over Z the model for $\Pr(Y|X)$ no longer takes the form of a logistic regression model.

For contrast, let's consider risk ratios estimated using a GLM with a log link. Under this model we have the conditional model

$$\Pr(Y|X,Z) = e^{\beta_0+\beta_X X+\beta_Z Z}$$

Supposing as above that Z is binary, we have

$$\begin{aligned}
\Pr(Y|X) &= e^{\beta_0+\beta_X X+\beta_Z \times 1}\Pr(Z=1) + e^{\beta_0+\beta_X X+\beta_Z \times 0}\Pr(Z=0) \\
&= e^{\beta_0+\beta_X X+\beta_Z} p_Z + e^{\beta_0+\beta_X X}(1-p_Z) \\
&= e^{\beta_0+\beta_X X+\beta_Z} p_Z + e^{\beta_0+\beta_X X}(1-p_Z) \\
&= e^{\beta_0+\beta_X X}\left\{\beta_Z p_Z + (1-p_Z)\right\} \\
&= e^{\beta_0^*+\beta_X X}
\end{aligned}$$

where $\beta_0^* = \beta_0 + \log\left\{e^{\beta_Z} p_Z + (1-p_Z)\right\}$. This shows that log risk ratio without adjusting for Z is β_X, the same value as in the adjusted analysis. Therefore the GLM with a log-link does possess the collapsibility property.

Discuss: What are the interpretations of the marginal and conditional risk differences, risk ratios and odds ratios.

See Practical 5 solutions.

Discuss: For a patient with a large kidney stone, which treatment effect estimate(s) are most relevant to inform their treatment?

For a patient with a large kidney stone, the conditional treatment effect (conditional on $Z = 1$) is most relevant. The marginal treatment effect estimates is of more relevance to a policy maker.

Part B: A simulated example with binary outcome

In this part we some simulated data (simdata_binary.dta) on a binary treatment X, binary outcome Y and binary covariate Z. The covariate Z is known to be measured temporally prior to X, meaning that we can be sure it does not lie on the causal pathway between X and Y.

4. Quantify the relationship between X and Z.

   There is no marginal association between X and Z. That is $\Pr(X = x|Z = z) = \Pr(X = x)$ (for both values of x and z).

5. Using logistic regression (using a saturated model) estimate the conditional probabilities $\Pr(Y = 1|X = x, Z = z)$ for $x = 0, 1$ and $z = 0, 1$, and hence the conditional risk differences

$$\Pr(Y = 1|X = 1, Z = z) - \Pr(Y = 1|X = 0, Z = z), \qquad z = 0, 1$$

   The estimated conditional risk difference in the $Z = 0$ group is 0.4, and it is also 0.4 in the $Z = 1$ group. There is no modification of the effect of X on Y by Z here (on the risk difference scale).

6. Without performing any further calculations, what do you expect the marginal risk difference to be?

   Because risk differences are collapsible, and because there is no modification of the effect of X on Y by Z, we expect the marginal risk difference to be equal to the two conditional risk differences, i.e. to be 0.4.

7. From your logistic regression in question 5, what are the estimates of the conditional odds ratios

$$\frac{\Pr(Y = 1|do(X = 1), Z = z)/\Pr(Y = 0|do(X = 1), Z = z)}{\Pr(Y = 1|do(X = 0), Z = z)/\Pr(Y = 0|do(X = 0), Z = z)}, \qquad z = 0, 1$$

   The conditional OR is 9, and this is given both values of Z. There is no interaction between X and Z.

8. Fit another logistic regression to obtain an estimate of the marginal odds ratio

$$\frac{\Pr(Y = 1|do(X = 1))/\Pr(Y = 0|do(X = 1))}{\Pr(Y = 1|do(X = 0))/\Pr(Y = 0|do(X = 0))}$$

   What assumption do you make? Explain the difference between this estimate and your estimates in question 7.

   In this example Z is not confounder, as we saw no association between X and Z. Therefore the marginal OR above can be estimated using a logistic regression of Y on X, i.e. we do not need to use standardization to estimate the probabilities $\Pr(Y = 1|do(X = x))$. Fitting the logistic regression of Y on X gives an OR estimate of 5.44. The marginal OR estimate is quite different from the conditional estimate (5.44 vs 9) - this is not due to confounding (there is no confounding here), but due to non-collapsibility.

Part C: A simulated example with binary outcome and continuous confounder

In this part we some simulated data (simdata_binary2.dta) on a binary treatment X, binary outcome Y and continuous variable Z, where Z confounds the association between X and Y.

9. Fit a logistic regression of Y on X and Z and their interaction. What is the conditional odds ratio for an individual with (a) Z equal to its median value in the data, (b) Z equal to its 10th percentile in the data, (c) Z equal to its 90th percentile in the data.

   The conditional OR estimates are (a) 0.93, (b) 1.49, (c) 0.58. Looking at the model output, there is quite strong evidence here of an interaction between X and Z - the coefficient is -0.06, with p-value 0.013. There is a qualitative interaction between X and Z in this example - for some values of Z the treatment is beneficial, and for other values it is detrimental.

10. Using the empirical standardization method introduced in session 5, obtain an estimate of the effect of X on Y using a marginal odds ratio.

    The marginal OR estimate is 0.87.

11. Compare this with the odds ratio from a regression of Y on X alone. What is the reason for the difference between this estimate and that in question 10?

    The crude marginal OR estimate from the regression of Y on X alone is 1.09. This differs from the estimate in question 10 (0.87) because the estimate in question 10 is adjusted for Z - i.e. we used standardization to adjust for Z and estimate the marginal probabilities $\Pr(Y = 1|do(X = x))$. The crude marginal OR estimate is a biased estimate of the causal marginal OR because it does not control for confounding by Z.

Discuss: What are the interpretations of the marginal and conditional odds ratios?

The (causal) marginal OR is 0.87. The interpretation is that if everyone in this population were to receive the treatment ($X = 1$) the estimated odds of $Y = 1$ would be 13% lower than had everyone not received the treatment ($X = 0$).

The conditional OR for an individual with Z at the median value in the population ($Z = 27.57$) is estimated to be 0.93. This means that if a person with this value of Z received the treatment their estimated odds of $Y = 1$ would be 7% lower than had they not received the treatment.

The conditional OR for an individual with Z at the 10th percentile value in the population ($Z = 20.03$) is estimated to be 1.49. This means that if a person with this value of Z received the treatment their estimated odds of $Y = 1$ would be 49% higher than had they not received the treatment.

The conditional OR for an individual with Z at the 90th percentile value in the population ($Z = 35.11$) is estimated to be 0.58. This means that if a person with this value of Z received the treatment their estimated odds of $Y = 1$ would be 42% lower than had they not received the treatment.