Scaled logit model with inclusion of unexposed population

Arseniy Khvorov

11/12/2019

All relevant code is at github.com/khvorov45/sclr-lowbase

1 Context

The aim is to investigate the ralationship between a covariate (e.g. antibody titre) and a binary outcome (e.g. infection status).

If the sample contains people unexposed to the pathogen (and therefore at no risk of the outcome), including them into the analysis will not bias the estimates under the scaled logit model

$$P(Y=1|E=1)=\frac{\lambda}{1+\exp(\beta_0+\beta_T T)}$$
$$P(Y=1|E=0)=0$$
$$P(Y=1)=\frac{P(E=1)\lambda}{1+\exp(\beta_0+\beta_T T)}=\frac{\lambda^*}{1+\exp(\beta_0+\beta_T T)}$$

Where $\lambda^*=\lambda P(E=1)$.

Including the unexposed into the analysis has the expected effect of lowering the baseline estimate by the probabiity of exposure. The infection curve can be expected to have a lower top plateau but the same logistic slope and intercept. The protection curve can be expected to be unaffected since it only depends on the β parameters.

However, including the unexposed into the analysis has an effect on the expected standard errors of the parameter estimates.

2 Effect of the unexposed on SE of parameter estimates

To investigate the effect of including the unexposed population into the analysis on the standard errors of the parameter estimates, I simulated 10,000 observations from the model

$$P(Y=1)=\frac{P(E=1)\exp(\theta)}{(1+\exp(\theta))(1+\exp(\beta_0+\beta_T T))}$$
$$\theta=\log(\frac{\lambda}{1-\lambda})$$
$$\lambda=0.5\quad\beta_0=-5\quad\beta_T=1.5$$

at different values of $P(E=1)$.

I then fit the scaled logit model using maximum likelihood to all data (general population, unexposed included) and the subset with just the exposed (unexposed excluded). The results of 10,000 simulations at each value of $P(E = 1)$ from 0.1 to 1 are in Figure 1.

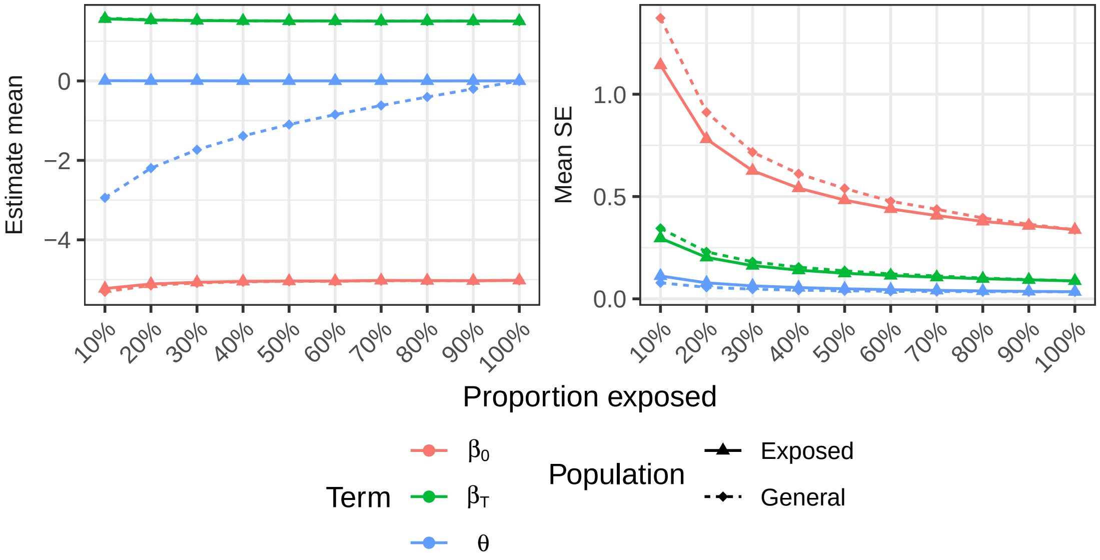

Figure 1: The results of 10,000 simulations at each parameter combination. Points represent values at which simulations were performed. Points and lines are colored based on the estimated term they belong to. The solid lines and triangles show estimated mean (left panel) and mean standard error (right panel) of estimates obtained from fitting models to the exposed population (i.e. unexposed excluded). The dashed lines and rhombi show the same for the general population (i.e. unexposed included).

As expected, including the unexposed into the analysos has the effect of lowering the estimated baseline probability but has no appreciable effect on the expected estimates of the other parameters.

Including the unexposed into the anlalysis also has the effect of increasing the expected standard errors of the β parameters (especially the intercept β_0) by 5-10% when the exposed proportion is less then 50%.

3 Using household infection as an indicator of exposure

One way to distinguish between the exposed and the unexposed populations is to use infection status within the household. If there is at least one infection in a household, all of its members are assumed to have been exposed. If there are no infections in a household, none of its members are assumed to have been exposed.

However, there are likely to be households with memebers who were exposed but where nobody got infected. These households would be excluded from the analysis which may create problems for the validity of the estimates.

To investigate this, I simulated 10,000 observations from the same model as in Section 2 except now everyone was assigned to a hoousehold. Each household had 4 subjects. The entire household was either exposed or unexposed according to the specified (expected) proportion. There was no random effect associated with being in a household. The results of 10,000 simulations at each value of proportion (of households) exposed are in Figure 2.

Using household infection as a measure of exposure appears to not introduce any notable bias in the estimates. The expected standard errors from this sample are similar to those that would have been obtained if the true infection status was known and everyone who was truly exposed was included into the analysis.

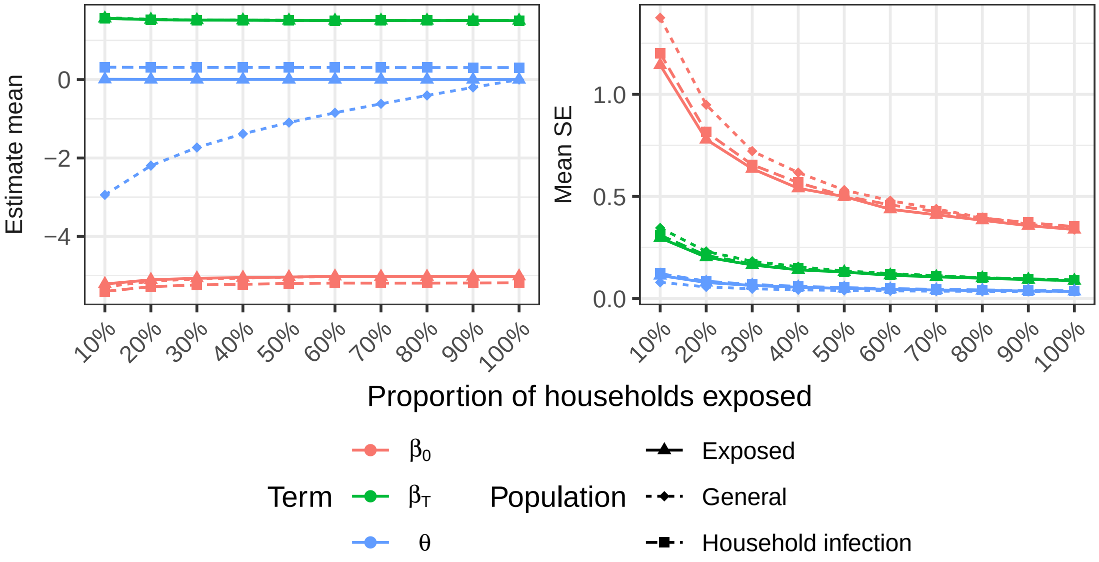

Figure 2: The results of 10,000 simulations at each parameter combination. Points represent values at which simulations were performed. Points and lines are colored based on the estimated term they belong to. The solid lines and triangles show estimated mean (left panel) and mean standard error (right panel) of estimates obtained from fitting models to the exposed population (i.e. unexposed excluded). The dashed lines and rhombi show the same for the general population (i.e. unexposed included). The dash-dotted lines and squares show the same for the population with at least one infection in the household.

4 Conclusion

Including the unexposed into the analysis offers no benefits and has the detriment of increasing estimate error (despite the fact that there is more data with the unexposed included). If there is a good way to isolate the unexposed subjects in the analysis, those observations should be excluded. Household infection appears to be a reasonable measure of exposure despite the fact that some exposed households will likely be excluded.