**Notes on regression models and estimated statistics**

For my second study, I will investigate the familial aggregation of persistence to methotrexate in monotherapy (mono-MTX) within early rheumatoid arthritis (RA). To do this, I test for an association between a family history of treatment persistence with treatment persistence in an index patient. More precisely, I pair up early RA patients treated with mono-MTX with their first-degree relatives that also had early RA and were similarly treated with mono-MTX. Persistence, in both the index patient and their relative(s), is then defined as a binary yes-no variable. Thus, we have categorical variables in a fairly straightforward data set. The question that remains is *what statistical model* do we use to test for an association, and *what statistic* should we estimate and present in our article?

The common, most obvious choice is of course logistic regression. Logistic regression allows us to fit a model, control for relevant covariates, produce estimates of the magnitude of association and ultimately test them for statistical significance [1 (Chapter 5)]. This model estimates odds ratios between binary variables, which in our example would mean that the acquired statistic would be the estimated *odds of remaining on treatment given a family history of treatment persistence*. Now, logistic regression is of course a valid model (given that the underlying assumptions are satisfied) but are odds ratios really the statistics that we are interested in? [2-8].

The issue with the odds ratio lies in its connection to the relative risk. In most epidemiological studies, the true statistic of interest is the relative risk, a concept more readily understood by non-statisticians as it describes the increased risk of those exposed compared to those not exposed. The issue lies within the fact that the odds ratio approximates the relative risk, when the outcome of interest is rare within the population. One can write the relationship as

where *πi*denotes the prevalence of outcome *i* within the population [1 (page 47)]. Therefore, if the outcome is rare, then the relative risk is closely approximated by the odds ratio; if it is common, the odds ratio will overestimate the risk. An added complexity lies in the fact that the above discussed prevalence is the prevalence within the population studied, meaning that for very rare outcomes it may be impossible to either collect data on the population or obtain a large enough sample to have sufficient data on the individuals with the rare outcome. This is of course mitigated by sample ascertainment as is done in the case-control setup, where cases (individuals with the outcome) are sampled and then paired with controls (individuals without the outcome). But then, due to the sample ascertainment, the above relationship no longer holds and only the odds ratio can be estimated using logistic regression.

Neither of the above complexities make logistic regression and the odds ratio inappropriate and it is important to remember that they are still valid for this type of data and analysis in general. However, it isn’t guaranteed that the results from such a model will be correctly interpreted when presented within a publication and readers (or worse, the authors) may *incorrectly* interpret the odds ratio as a close approximation of the relative risk [2, 5, 6]. In the context of our study, the prevalence *πi* is also knowingly *not small* (i.e., the outcome is not rare). In fact, we know that our sample (individuals within the Swedish Rheumatology Quality Register) has near complete coverage of the total population (Swedish early RA patients) [9], so the proportion of individuals who are persistent to mono-MTX within our data should be a very good estimate of the prevalence *πi*. For us, those proportions lie at about 0.7 for persistence at one year and 0.5 for persistence at three years, which will lead to odds ratios that *do not* approximate the underlying relative risk very well. The important question here is not to be posed to us in the form of statisticians but as general researchers: if we can produce research that is not prone to misinterpretation, is it not important that we go out of our way to do so? I would argue that indeed, why present odds ratios if we can acquire the relative risk estimates?

Various adjustments exist through which we can transform the estimated odds ratio obtained from a logistic regression model to an estimate of the relative risk, with the most commonly discussed being the one described by Zhang and Yu [10]. While well established, this approach has received critiques [11] and I will thus not discuss it here, instead choosing to focus on two regression models that are, to my knowledge, the currently best available ones for this type of situation: the log-binomial regression [11-15] and the robust Poisson regression [11, 16].

In simplified terms we can think of log-binomial regression as an alternative to logistic regression where the link function is changed from the logit to the log (in R, this would correspond to changing the `family` parameter of the `glm()` function from `family = binomial(link = ‘logit’)` to `family = binomial(link = ‘log’)`). The obtained estimate is then the *log relative risk* instead of the standard log odds ratio. We can see this with a simplified example for the case of the 2x2 contingency table. For a given link function *g*, the GLM is given by

where *x* is of course binary and *π(x)* denotes the probability of *x* [1 (page 124)]. Now if we insert values *0* and *1* into the equation above we can extract the value of *β* which is given by

Now if the link function *g* is simply the identity link, then *β* corresponds to the difference of proportions. If instead we use a log link function, *g(x) = log(x)*, then we get

Likewise, if we use the logit as link function we will arrive at the log odds ratio as expected. This is of course a highly simplified example but I find it useful to give here as it shows a quite intuitive connection between logistic regression and log-binomial regression which may make the concept of the latter less intimidating for those accustomed to using the former.

Now, before moving on to a discussion of the robust Poisson regression, one needs to cover the limitations of using the log-binomial regression. First and foremost, log-binomial regression often suffers from convergence issues [11-14]. Maximum likelihood estimation of the model parameters is commonly done through iteratively reweighted least squares, an algorithm that generally converges poorly for this type of data as a result of the link function not being the canonical link function. Secondly, log-binomial regression has been shown to produce more liberal (i.e., narrower) confidence intervals than what is expected, which could lead to the incorrect interpretation of borderline significant results [11].

A second alternative exists in the form of the robust Poisson regression. I will avoid discussing the theory behind the model (since I find it less intuitive and obvious than the log-binomial regression) but instead refer the interested reader elsewhere [11, 16]. What is worth noting is that this model relies on assuming a Poisson model for the binomial data which thus leads to misspecified error terms, meaning that confidence intervals will generally be more conservative than necessary [11, 16]. This can however be accounted for by employing a sandwich estimator through which ‘robust’ standard errors can be computed [16-18].

It is worth mentioning that both methods have been shown to produce generally similar, highly concordant results, meaning that the choice between them is mainly a subjective question of taste and preference [11, 13, 19, 20]. A recent simulation study did indicate that robust Poisson regression would be more stable under model misspecification, though the authors did not find their results to suggest that the difference would be meaningful [20] Personally, I prefer the log-binomial regression for our data since we already have an issue with model misspecification leading to more liberal confidence intervals as a result of family clusters within data (i.e., data being repeated for individuals with multiple first-degree relatives). As such, we already rely on using robust standard errors to obtain appropriate confidence intervals, something that I am unsure how it will affect the need for robust standard errors for the Poisson model. Additionally, as I find the theory behind the model less intuitive for robust Poisson regression, it seems ‘safer’ to rely on the method I have a better grasp of.

As stated earlier, log-binomial regression suffers from potential convergence issues, something that increases with small sample sizes and continuous covariates to name two things [11-14]. Various alternatives exist to mitigate this issue, from binning continuous covariates to relying on more sophisticated estimation algorithms [12, 21-23]. Many of these alternative estimation algorithms are implemented within the ‘logbin’ package for R which I would recommend for practical applications [12].

**References**

[1] A. Agresti. *Categorical data analysis.* Hoboken; Wiley; 2002.

[2] D. Altman, J. Deeks and D. Sackett. Odds ratios should be avoided when events are common [letter]. *BMJ,* 317:1318, 1998.

[3] M. Bracken and J. Sinclair. When can odds ratios mislead? Avoidable systematic error in estimating treatment effects must not be tolerated [letter; comment]. *BMJ*, 317:1156, 1998.

[4] H. Davies, I. Crombie and M. Tavakoli. When can odds ratios mislead? *BMJ*, 316: 989-991, 1998.

[5] J. Deeks. When can odds ratios mislead? Odds ratios should be used only in case-control studies and logistic regression analysis [letter; comment]. *BMJ*, 317:1155-1156, 1198.

[6] W. Holcomb, T. Chaiworapongsa, D. Luke and K. Burgdorf. An odd measure of risk: use and misuse of the odds ratio. *American College of Obstetricians and Gynecologists,* 98(4):685-688, 2001.

[7] D. Sackett, J. Deeks and D. Altman. Down with odds ratios. *Evid. Based. Med.,* 1:164-166, 1996.

[8] D. Taegar, Y. Sun and K. Straif. On the use, misuse and interpretation of odds ratios. *eBMJ*, 1998. DOI: <https://doi.org/10.1136/bmj.316.7136.989>.

[9] SRQ, 2021. Visualization and analysis platform (VAP) [Accessed 27 Apr 2021].

[10] J. Zhang and K. Yu. What’s the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA,* 280:1690-1691, 1998.

[11] L. McNutt, C. Wu, X. Xu and J. Hafner. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am. J. Epid.,* 157:940-943, 2003.

[12] M. Donoghoe and I. Marschner. **logbin**: An R package for relative risk regression using the log-binomial model. *J. Stat. Softw.,* 86(9), 2018.

[13] T. Skov, J. Deddens, M. Petersen and L. Endahl. Prevalence proportion ratios: estimation and hypothesis testing. *Int. J. Epid.,* 27:91-95, 1998.

[14] S. Wacholder. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am. J. Epid.,* 123(1):174-184, 1986.

[15] S. Wallenstein and C. Bodian. Inferences on odds ratios, relative risks and risk differences based on standard regression programs. *Am. J. Epid.,* 126(2):346-355, 1987.

[16] G. Zou. A modified Poisson regression approach to prospective studies with binary data. *Am. J. Epid.,* 159(7):702-706, 2004.

[17] D. Lin and L. Wei. The robust inference for the Cox proportional hazards model. *J. Am. Stat. Assoc.,* 84:1074-1078, 1989.

[18] R. Royall. Model robust confidence intervals using maximum likelihood estimators. *Int. Stat. Rev.,* 54:221-226, 1986.

[19] A. Barros and V. Hirakata. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Med. Res. Methodol.,* 21, 2003.

[20] W. Chen, L. Qian, J. Shi and M. Franklin. Comparing performance between log-binomial regression and robust Poisson regression models for estimating risk ratios under model misspecification. *BMC Med. Res. Methodol.,* 18, 2018.

[21] J. Deddens, M. Petersen and X. Lei. Estimation of prevalence ratios when PROC GENMOD does not converge. In *Proceedings of the 28th Annual SAS Users Group International Conference,* paper 270-280, 2003.

[22] I. Marschner. **glm2**: fitting generalized linear models with convergence problems. *The R Journal*, 3(2):12-15, 2011.

[23] I. Marschner. Combinatorial EM algorithms. *Stat. Comput.,* 24(6):921-940, 2014.