Mathematical Statistics
Stockholm University
Master Thesis **2015:6**
http://www.math.su.se

Comparison of the Rank-Ordered Logit and Between-Within Regression Models

Marielle Andersson*

September 2015

Abstract

When we have epidemiological data and want to investigate the association between an outcome and an explanatory variable we need to adjust for potential confounders that otherwise can cause a statistically significant association between these factors. In this thesis we compare two regression models that can be applied in order to adjust for confounders when the outcome is continuous by dividing the population into clusters, namely the between-within and the rank-ordered logit model. The between-within model is an extension of the generalized linear model where we include the cluster specific mean of the explanatory variable as an additional covariate. We thereby divide the regression into a within- and a between-effect, where the within-effect is not affected by the confounders shared within a cluster. The rank-ordered logit model assumes that the unknown information has a so called extreme value type I distribution and ranks the outcomes within a cluster. The resulting log likelihood function is equivalent to the likelihood of a stratified Cox proportional hazards model, where all shared confounders within a cluster are matched away. We compare these two models in a simulation study and also apply them on two datasets. The first dataset contains information about blood glucose measurement from the National University Hospital in Singapore and we study the association between the variation in blood glucose level and the mean daily measurement frequency, and we conclude that there is a statistically significant positive association. The second dataset is from the Karolinska mammography project for risk prediction of breast cancer (KAR, 2015), where we analyse risk factors for mammographic density. We focus on post menopausal women and compare the results from the analysis of pairs of unrelated women and pairs of sisters. We find that it is important to adjust for age and body mass index, but not important to adjust for confounding by shared childhood environment and genetic factors. We conclude that whether or not a woman has had hormone replacement therapy or a history of benign breast disease are associated with percent dense volume which is a measure of mammographic density. Also, the age at first birth is associated with mammographic density. We conclude that the between-within model is preferable to the rank-ordered logit model. The advantage of these two models is that we can adjust for unmeasurable confounders. However, both models are biased if the confounders are not completely shared within a cluster and should therefore be applied with caution.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: maan0129@gmail.com. Supervisor: Pieter Trapman.