Electronic Supplementary Material 2

Minimum sample size determination for coefficients in DCEs

A key information requirement in order to determine the minimum sample size is the information that is needed to determine the precision of the estimates of interest, given a specific sample size. Below we first provide details on constructing an estimate of the precision of the estimated model parameters. We then outline the steps needed to infer the minimum sample size, given the desired confidence and power levels.

In choice models, relying on the asymptotic theory of maximum likelihood estimation [1], the limiting distribution of the estimated coefficients γ is given by

$$\sqrt{N}(\hat{\gamma} - \gamma) \sim N\left(0, \Sigma_{\gamma}\right).$$

This equation describes the distribution of the parameter estimates relative to the true value for sufficiently large values of N. The properties of this distribution are often used to facilitate hypothesis testing on the model parameters γ. The variance-covariance matrix $\boldsymbol{\Sigma}_{\gamma}$ of this normal distribution equals the inverse of the Fisher information matrix [1] for the specific DCE design and choice model specification. Details of this variance-covariance matrix for various models have been provided elsewhere in the literature, in particular in the literature on generating efficient designs, see e.g. Kessels et al [2], Sandor & Wedel [3], and Bliemer & Rose [4]. This appendix provides the details for the multinomial logit (MNL) model, but the analysis extends to any other model as long as a specification of $\boldsymbol{\Sigma}_{\gamma}$ can be obtained.

Let the DCE under consideration be defined by S choice sets with each choice set containing J profiles or alternatives. The set of attribute values corresponding to the j^{th} alternative of the s^{th} choice set is given by x_{sj} and the corresponding utility assigned to this alternative is given by

$$U_{sj} = x'_{sj}\gamma + \varepsilon_{sj}$$

The MNL model is obtained by assuming an extreme value type I distribution of the unobserved component ε_{sj} in the random utility specification. The Fisher information matrix (i.e. the inverse of the variance covariance-matrix) for the MNL of this DCE design when N independent responses have been obtained is given by

$$I(X,\gamma) = N\sum_{s=1}^{S} X'_s(P_s - p_s p'_s)X_s = (\Sigma_\gamma)^{-1}. \qquad \text{(Eq. A.1)}$$

Here X_s stacks the attributes x_{sj} for each of the J alternatives in choice set s, p_s stacks the J choice probabilities for the alternatives and P_s is a matrix of zeros and p_s on its diagonal. This expression for Σ_γ depends on the value of γ itself, through the choice probabilities in p_s, which are given by

$$p_{sj} = \exp(x'_{sj}\gamma) \,/\, \sum_{j=1}^{J} \exp(x'_{sj}\gamma).$$

So far, we have outlined in detail how to obtain an estimate of $\Sigma_{\gamma k}$, which will be used in hypothesis testing. We now want to assess whether a parameter γ_k is equal to a hypothesized value γ_k^*. Given that there is always sampling error and we potentially get a very rare sample, we have to be satisfied knowing that a certain sample size will enable us to find a significant deviation from γ_k^* when testing at an 1-α confidence level in at least β·100% of the cases if the true γ_k deviates from γ_k^* by δ or more.

To test whether the estimated parameter value $\widehat{\gamma_k}$ differs from γ_k^*, we use the *z*-statistic defined as

$$z = (\hat{\gamma}_k - \gamma_k^*)/\sqrt{\Sigma_{\gamma k}/N}$$

With $\mathbf{z_{1-\alpha}}$ denoting the critical value for testing at the 1-α confidence level, a statistically significant difference will be found when the test statistic exceeds this value, so whenever

$$(\hat{\gamma}_k - \gamma_k^*)/\sqrt{\Sigma_{\gamma k}/N} > z_{1-\alpha}.$$

When the true coefficient γ_k deviates from the hypothesized value γ_k^* by $\delta>0$ or more, and we want to find a significant difference with at least a probability 1-β, we need a sample size such that

$$1-\beta < P\left\{(\hat{\gamma}_k - \gamma_k^*)/\sqrt{\Sigma_{\gamma k}/N} > z_{1-\alpha}\right\} =$$

$$P\left\{(\hat{\gamma}_k - \gamma_k - (\gamma_k^* - \gamma_k))/\sqrt{\Sigma_{\gamma k}/N} > z_{1-\alpha}\right\} =$$

$$P\left\{(\delta - (\gamma_k^* - \gamma_k))/\sqrt{\Sigma_{\gamma k}/N} > z_{1-\alpha}\right\} =$$

$$P\left\{-(\gamma_k^* - \gamma_k)/\sqrt{\Sigma_{\gamma k}/N} > z_{1-\alpha} - \delta/\sqrt{\Sigma_{\gamma k}/N}\right\} =$$

$$1 - \Phi\left(z_{1-\alpha} - \delta/\sqrt{\Sigma_{\gamma k}/N}\right) = \Phi\left(-z_{1-\alpha} + \sqrt{N}\delta/\sqrt{\Sigma_{\gamma k}}\right)$$

Intuitively, this result makes sense. For a given difference δ>0, we find that an increasing N increases the probability of finding a significant deviation and this also holds for increases in δ. A larger variance ($\Sigma_{\gamma k}$) induces more noise in the estimated coefficient, lowering the probability of finding a significant effect, which is also the case for testing at a higher confidence level α.

Applying the inverse cumulative distribution function (CDF) of the normal distribution, i.e. applying Φ^{-1} on the left and right, we obtain:

$$\Phi^{-1}(1-\beta) = z_{1-\beta} < -z_{1-\alpha} + \sqrt{N}\delta / \sqrt{\Sigma_{\gamma k}}$$

We can now isolate $\sqrt{N}$ and take squares on both sides to obtain the required sample size given by

$$N > \left((z_{1-\beta} + z_{1-\alpha})\sqrt{\Sigma_{\gamma k}} / \delta\right)^2$$

In the main text, we have implicitly used as hypothesized value zero, indicating the absence of an influence of an attribute level. This also implies that the coefficient values γ correspond to the effect sizes δ in the equations above. There is, however, no need to use the same parameter values as effect sizes and to obtain an estimate of the variance covariance matrix $\boldsymbol{\Sigma}_{\gamma\mathbf{k}}$.

The required sample size for MNL coefficients in a DCE can be calculated given

- The confidence level α at which tests are performed
- Information to obtain $\Sigma_{\gamma k}$, which requires knowledge on the DCE design and an initial guess of all the model coefficients. In the case of MNL this is the set of attribute weights γ
- The minimum probability $1-\beta$ that one wants to detect a deviation of a regression coefficient from a hypothesized value, given that the deviation is at least δ

The required sample size is then given by $N > \left(\left(z_{1-\beta} + z_{1-\alpha}\right)\sqrt{\Sigma_{\gamma k}} / \delta\right)^2$

References

1. Amemiya T. Advanced econometrics; 1985.
2. Kessels R, Roos P, Vandebroek M. A comparison of criteria to design efficient choice experiments. Journal of Marketing Research. 2006;43(3):409-19.
3. Sandor Z, Wedel M. Profile Construction in Experimental Choice Designs for Mixed Logit Models. Marketing Science. 2002;21(4):455-75.
4. Bliemer MCJ, Rose JM. Construction of experimental designs for mixed logit models allowing for correlation across choice observations. Transportation Research Part B: Methodological. 2010;44(6):720-34.