University of Warwick, Department of Sociology, 2012/13

SO201: SSAASS/Surveys and Statistics (Richard Lampard)

**Week 12: Results from a thematic examination of examples of published multivariate analyses (using linear regression and logistic regression)**

**Choice of technique**

Gauthier: Interval-level scale (constructed from *z*-scores), hence uses OLS linear regression

Weinberger *et al*.: Divorce is a dichotomous outcome (two categories), so the use of logistic regression is OK, although if they had durations from marriage to divorce they could have used an event history technique (see below).

Chatzitheochari and Arber: The reduction of an interval-level dependent variable (hours of sleep) to give a binary variable that is then used within a logistic regression analysis seems contentious/open to criticism (although it may reflect the imprecision of the original measure: see p35).

**Operationalisation of concepts (in relation to theory and in terms of detail)**

Gauthier: While her operationalization of concepts generally seems sound, her ‘minority status’ measure is very crude (two categories only), presumably because of small sub-sample sizes for more specific minority ethnic groups?

Gauthier: Uses household income per equivalent adult, which is preferable to overall household income.

Chatzitheochari and Arber: The operationalization of their measures is described in detail and generally seems sensible (with the exception of their dependent variable: see above).

Pampel: Notes problems with using self-reported smoking as an indicator of smoking behaviour.

**Impact of data source on results (generalisability; impact of sample design)**

Gauthier: Uses the fourth wave of a longitudinal study (BHPS), with the consequent unrepresentativeness that may arise because of attrition.

Gauthier: Reports using households rather than children as her units of analysis in her OLS linear regression, to avoid the impact of clustering of children within households on the precision of (estimated standard errors = quantities of sampling error for) her results, but does not report taking account of the impact on the precision of her results of the multi-stage sample design of her data source.

Weinberger *et al*.: Their sample is less than 50% of the original sample of a study based on University of Rochester students (limiting both the represesentativeness and the generalizability of the results?)

Chatzitheochari and Arber: Their data source is a diary-based study, with a low response rate (45%) BUT they report that the pattern of non-response does not appear to be differential, and they report weighting the results to correct for non-response.

Lee *et al*.: Note that (and explain why) their sample of pre-natal care service users in Philadelphia may not be representative of all Philadelphia mothers.

Lee *et al*.: Report correcting for the impact of the clustering of their respondents within health centers on the standard errors (precision) of their estimates.

Pampel: Notes that he needed to take account of changes between years in the nature of the survey data collected by his data source.

Pampel: Reports correcting for the impact of clustering, etc., within his data source’s sample design on the precision of the results (p531), and reports weighting the data to correct for a number of sources of unrepresentativeness.

**Inclusion of controls (including its value to the multivariate analysis) and omitted variables**

Gauthier: Includes very few controls, so the impact of things which might be correlated with income and explain its effects is not really taken into account.

Chatzitheochari and Arber: Include a reasonable number of potentially pertinent controls that might underpin differences in sleep according to working hours (e.g. age, social class, things relating more specifically to working hours).

Lee *et al*.: Include lots of pertinent-looking controls, and explicitly discuss some that they were unable to include (p1269).

**Assumptions of technique (e.g. linearity): plausibility and evidence of checks**

Gauthier: Uses log of income, apparently to increase the correlation (p251) with her dependent variable (deprivation index), presumably because the relationship with income is non-linear (which it still appears to be, even after income is log-transformed! (Fig. 2: p253)).

Gauthier: Fig. 2 (p253) suggests that there may be heteroscedasticity, but Gauthier does not comment on this, and does not report whether or not the residuals from her OLS regression appear to have a normal distribution.

Pampel: Uses quadratic effects to allow for non-linearity in the impact of year of survey (i.e. time).

**Presentation of results (Complete? Clarity?)**

Gauthier: Presents beta values as well as B values, allowing the effects of different variables to be compared in a meaningful way (in terms of strength).

Lee *et al*.: Present bivariate odds ratios, i.e. the overall effects for their independent variables, which can then be compared with the net effects that can be seen in their multivariate analyses.

Pampel: Presents some useful graphical displays to illustrate interaction effects, i.e. the changing impact of education over time, but does not provide the corresponding statistical testing results when he goes on to break this down according to ethnic group.

**Interpretation of results (magnitude, statistical significance and meaning of effects (B’s/Exp(B)’s))/discussion of statistical results in general**

Gauthier: Interprets her results sensibly, including in terms of their statistical significance.

Gauthier: Notes the relatively high level of variance explained by her OLS linear regression (39.6%)

Weinberger *et al*.: Include non-significant effects of variables, in a questionable way. In particular, their age at marriage ‘control’ variable is non-significant, which rings alarm bells (given that age at marriage has repeatedly been shown in the literature to have an impact on the risk of the divorce).

**Scope for Type I and/or Type II errors**

Weinberger *et al*.: The only significant effect has a *p*-value of 0.04, giving quite a substantial chance of this being a Type I error, given the overall number of tests carried out.

**Choice of reference categories for categorical variables**

Gauthier: Makes sensible comparisons between other family types and two-parent families, her reference category (which is a rather conventional approach, but conceptually meaningful and means the reference category is appropriately large).

Chatzitheochari and Arber: The reference category in their examination of an interaction appears poorly chosen, as it is the only category which appears to be distinctive in this context.

**Examination (or lack of examination) of interaction effects**

Gauthier: Does not report checking for any interaction effects (e.g. what about the possibility that family-type effects vary according to income level?)

Chatzitheochari and Arber: Report checking systematically for interaction effects between their independent variables, but present and interpret the sole significant interaction ineffectively (see above; note that the relevant part of the text focuses more on the main gender effect than on the interaction with class).

Lee *et al*.: Report checking systematically for interaction effects between their independent variables.

Pampel: Focuses on interaction effects relating to the impact of education and the impact of year of survey.

**Reasons for/value of presenting results from a series of models/regressions**

Weinberger *et al*.: The value of including the results from a series of nested logistic regressions is not evident.

Chatzitheochari and Arber: Use nested models to account for the impact of controls (other variables) on the impact of long-hours of work; this approach reveals a suppressed effect of long hours for men.

Lee *et al*.: Use a series of nested models to show how ethnic group effects change when control variables are included, and are accordingly able to explain much of the overall effect for one ethnic group and reveal a significant, previously suppressed effect for another ethnic group.

Pampel: Uses nested models to show how trends in the impact of education on smoking behaviour change when other variables are taken into account.

**Use of additional and/or more sophisticated techniques, and reasons for this…**

Gauthier: Uses principal components analysis to ‘unpack’ the underlying dimensions of her multi-item deprivation scale.

Lee *et al*.: Use Cox’s proportional hazards model to look at variations in breastfeeding duration.