Review of the manuscript 'systemfit: A package to estimate simultaneous equation systems in R'

This manuscript describes a new R package systemfit for inference in multi-equation models. The main parts are section 2 and 4. Section 2 gives the statistical background for the analysis tools available in systemfit while section 4 describes how to apply systemfit. The theory is adequately explained and the syntax in systemfit is nice and logical allowing complicated models to be quite easily fitted.

My only real concern is that systemfit relies on classic econometric (least squares) developed 40 years ago. Since then computational power has increased dramatically and superior methods may be available. First of all, full likelihood methods are available in existing software for structural equation modelling. In particular, R has a package called SEM. I am not sure what is gained with systemfit compared to this. Also it is a weakness that systemfit only allows analysis of complete cases.

Specific comments

Instrumental variable estimation has received much attention lately in biostatistics. Maybe it would help 'sell' systemfit if this was emphasized at bit more. For instance, instrumental variables could be mentioned directly in the introduction and included as a key word.

Section 2 gives a brief overview of the statistical background. This may be a bit hard to follow for people outside econometrics. It would help to include more references, especially in section 2.1.

Page 2: 'observations' should be explained earlier together with the presentation of the data structure. The number of observations should be stated.

Page 2: It should be clarified whether y-variables in one equation can be x-variables in another.

Page 4, line 82: 'may' should be 'must'. Here the difference is extremely important. Maybe the text could emphasize more clearly what is required of a variable if it is to be used as an instrumental variable. What is a good instrumental variable and what happens if the requirements are not fulfilled?

The difference between equation 20 and 23 is so smaller that maybe it should be noted in the text what it is.

Line 193-194: Language is confusing. It should be explained that calculation of some of the estimators earlier introduced require Sigma to be known. Often Sigma is not known so an estimator is used instead. However, then statistical properties may changed compared to the theory described earlier.

Line 196: It should be explained what is meant by a first step OLS.

Section 2.3: Somewhere in this section or (2.1) it should be noted that the efficiency results of section 2.1 depend on Sigma being known. And that a full maximum likelihood analysis may be superior.

Line 215-219: Alternative approach should be explained more clearly.

Line 375-379: Syntax on programming of instrumental variables should be made more clear. What is a one-sided formula? Does the code given imply that the sum of income, farmPrice and trend was used as and instrumental variable? If so, it should be stated in the text.

Section 2.7: Given the theory of the previous sections, I was confused about the Hausman test. Under the alternative hypothesis, independent variables are correlated with error terms. In this section it is stated that this leads to inconsistency of 3SLS. However, on page 5 (top) the opposite is stated? Would it not make more sense to compare SUR to 3SLS?

Smaller corrections:

Sometimes ':' is used just before an equation other times not.

Page 7, line 147: change 'doesn't' to 'does not'.

Line 135: '\beta^*' should be '\widehat{\beta^*}'

Line 209: first \widehat{\Sigma} should be \Sigma

Equation 63: '\Sigma' should be '\widehat{\Sigma}

Equation 65: Why two hats?

Line 299: 'interest' should be 'interested'

Page 14: 'dependant' should be 'dependent'