A Macroscopically Estimating Model with Endogeneity considered for Lane-mean Speeds and Speed Deviations (Proposal)

Background

Some transportation data are best modeled by a system of structural equations, including average vehicle speeds by lane and speeds deviations. In multi-lane traffic, the interrelation between lane mean vehicle speed and the speeds in adjacent lanes cannot be omitted as it influences a lot to the estimation of lane-mean speeds. And while modeling lane-speed deviations, V. Shankar and F. Mannering (1998) reveals that both adjacent lane speed deviations and adjacent lane speeds are very important. So when estimating the mean speeds of lanes in a multi-lane road, adjacent speeds serve as independent variables in the equations to build up a precise model. Similarly, when it turns to speed deviations, both speeds and speed deviations in adjacent lanes are included in independent variables dataset.

However, a latently serious estimation issue arises in the interrelated systems of equations if ordinary least squares (OLS) is employed to estimate equations separately and their interrelated structure is not considered. Because the uncorrelatedness of regressors and disturbances assumption of OLS is violated since one or more independent variables are endogenous and therefore leading to biased and incorrect estimation results.

For example, in an interrelated system of equations for estimating lane mean speeds in two-lane traffic, the model by OLS is like the following linear form (the subscript for observation number is not shown),

$$u_1 = \boldsymbol{\alpha_1} \cdot \boldsymbol{Z_1} + \lambda_1 \cdot u_2 + \varepsilon_1 \quad (1)$$

$$u_2 = \boldsymbol{\alpha_2} \cdot \boldsymbol{Z_2} + \lambda_2 \cdot u_1 + \varepsilon_2 \quad (2)$$

where u_i is the lane-mean speed of lane i . $\boldsymbol{Z_i}$ is the vector of the variables of traffic flow state, environment situations, etc. $\boldsymbol{\alpha_i}$, λ_i are the parameters waiting for being estimated. ε_i is disturbance term. i = 1, 2.

If equation (1) and (2) are estimated using OLS separately, the results are biased and incorrect since they violate one of the key assumptions of OLS which is known as uncorrelatedness of regressors and disturbances assumption mentioned above. Because the changes of dependent variables in the two equations (u_1 in (1), u_2 in (2)) discussed here will influence the value of the independent variables (u_2 in (1) and u_1 in (2)). That means u_2 in (1) or u_1 in (2) is not fixed in repeated samples. So the independent variable u_2 in (1) and u_1 in (2) are labeled as endogenous variables in (1) and (2), respectively. To summarise, using OLS to estimate (1) and (2) separately will produce biased and inconsistent estimates of model parameters.

In *V. Shankar and F. Mannering (1998)*, a structural equations approach is used to model the lane mean speeds and lane speed deviations where three-stage least squares (3SLS) is applied to handle the endogeneity problem. It executes an empirical study based on this approach in which the relevant data of a six-lane road (three lanes each direction) are loaded including environmental data, temporal data and traffic flow factors and achieves considerable results. However, although it copes with the endogeneity of lane mean speeds and speed deviations giving rise to best linear unbiased estimators (BLUE), it is discussed in very wide time windows (1 hour) and a long road segment that may make it less possible to be applied in practice.

In addition, *Wen Cheng et al. (2018)* propose hierarchical models based on Bayesian framework to address the same problem. However, in their empirical study, the influence of trucks in traffic flow doesn’t be considered in their models which may result in unrealistic and unreliable output. And the models are employed in the estimations of mean speeds and speed deviations in two-lane traffic which is not representative. Besides, the interpretation of how the Bayesian framework handles the endogeneity of speeds and speed deviations is not given.

To summarize, a more integrated model is needed to model and address the endogeneity of the lane-mean speeds and lane-speed deviations. And then based on a better understanding of the mean speeds and speed deviations on lanes, more studies can be raised to address the safety problems of road traffic, the design problems of roads and so on. So it is meaningful and necessary to follow up this research.

Model Constructing

This research proposes a structural system of equations to predict the lane-mean speeds and lane-speed deviations in which the endogeneity of variables is considered. In order to measure the influence of the forward traffic to the traffic upstream in the same lane, a road segment is divided into several smaller segments and multi-lane kinematic wave (KW) model (*J. A. Laval and C. F. Daganzo (2004)*) is employed to estimate the traffic flow state (traffic speed, traffic volume, etc.) in each cell. And then based on the estimation results of multi-lane KW model, interrelated equations are applied in each cell to estimate model parameters for the purpose of predicting speeds and speed deviations in next time window of each cell where 3SLS is also used to address the problem of endogenous independent variables. According to *V. Shankar and F. Mannering (1998),* lane-mean speeds are determined by not only environmental data, temporal data and traffic flow data, but also the vehicle speeds in adjacent lanes. And lane-speed deviations are related to their basic data (environmental data, temporal data and traffic flow data), lane speed deviations and also lane mean speeds in adjacent lanes. Turns to the model constructed in this research, lane-mean speeds are related to their basic data, vehicle speeds in adjacent cells and vehicle data in the next cell in forward, while lane-speed deviations are influenced by their basic data, lane-mean speeds and lane-speed deviations in adjacent cells and nearest cell in forward.

As a result, for lane-mean speeds, the equation system can be written as below (the subscript for the segment is not shown),

$$u_i = \beta_i + \boldsymbol{\alpha}_i \cdot \boldsymbol{Z}_i + \boldsymbol{\lambda}_i \cdot \bar{\boldsymbol{u}}_i + \varepsilon_i \quad (3)$$

where i is the lane number. u_i is the lane mean speed. $\boldsymbol{Z}_i$ is the vector of exogenous variables that affect lane-mean speed. $\bar{\boldsymbol{u}}_i$ is the vector of endogenous variables including mean speeds in the lanes adjacent to lane i and mean speed of next segment in lane i. $\beta_i, \boldsymbol{\alpha}_i, \boldsymbol{\lambda}_i$ are estimable coefficients. ε_i is disturbance term.

Similarly, the equation system for lane-speed deviations can be written as below (subscript for segment is not shown),

$$\sigma_i = \rho_i + \boldsymbol{\eta}_i \cdot \boldsymbol{V}_i + \boldsymbol{\omega}_i \cdot \bar{\boldsymbol{\sigma}}_i + \nu_i \quad (4)$$

where i is the lane number. σ_i is the standard deviation of speed in lane i. $\boldsymbol{V}_i$ is the vector of exogenous variables that affect lane-speed deviation. $\bar{\boldsymbol{\sigma}}_i$ is the vector of endogenous variables including mean speeds, speed deviations in the lanes adjacent to lane i and next segment in lane i. $\rho_i, \boldsymbol{\eta}_i, \boldsymbol{\omega}_i$ are estimable coefficients. ν_i is disturbance term.

In order to obtain unbiased and consistent results in estimating equation (3) and equation (4) both of which contain endogenous independent variables, 3SLS is appropriate. The executing procedure can be seen as follow,

Stage 1: Getting two-stage least squares (2SLS) estimates of the model system;

Stage 2: Applying the estimates from Stage 1 to calculate the disturbances of the system;

Stage 3: Generalized least squares is used to compute coefficients.

And the procedure of 2SLS is shown below,
Stage 1: Endogenous variables regressed against exogenous variables;
Stage 2: Treating the results of Stage 1 as instruments and then estimating each equation by OLS.
Attributed to some standard statistical software packages which contain 3SLS, it is not a trouble to apply 3SLS to analyst relevant problem anymore.

Empirical Study

An empirical test is researched rely on the data from Caltrans Performance Measurement System (PeMS). Based on this empirical study, a comparison between the results of the model constructed in this research and the model proposed in *V. Shankar and F. Mannering (1998)* will be given. And an analysis of superiority and shortcomings is expected.