Random Coefficient Regression

Bengt Muthén, Linda Muthén & Tihomir Asparouhov

April 27, 2015

The linear regression model assumes that all individuals come from a population with a single slope β. This assumption can be relaxed by allowing the slope to vary across individuals and be predicted by other covariates. The varying slope is referred to as a random slope, an unobserved continuous variable β_i. Models of this kind are also called random coefficient models and have been discussed in Hildreth and Houck (1968), Johnston (1984), Swamy (1971), and Weisberg (2014). Consider a model where a random slope is predicted by a covariate z which moderates the effect of a covariate x on a dependent variable y. Random coefficient modeling expresses this as two linear regression equations, one with y as the dependent variable and one with the random slope β_{1i} as the dependent variable,

$$y_i = \alpha + \beta_{1i}\, x_i + \beta_2\, z_i + \epsilon_i, \tag{1}$$

$$\beta_{1i} = \beta_0 + \beta_1\, z_i + \delta_i. \tag{2}$$

The residuals ϵ and δ are allowed to covary. The model can be compared to regular regression with an interaction between the covariates x and z by inserting (2) into (1),

$$y_i = \alpha + \beta_0\, x_i + \beta_1\, z_i\, x_i + \delta_i\, x_i + \beta_2\, z_i + \epsilon_i. \tag{3}$$

The product term $\beta_1 z_i x_i$ is present also in regular regression with an interaction, but random coefficient regression adds the term $\delta_i x_i$. If the δ residual is zero the two models are the same. The random coefficient model includes a non-zero δ residual to acknowledge that z may not be the only moderator and that the R^2 for

β_{1i} in the regression equation (2) is more realistically less than one. The model has two more parameters than the regular regression model, the variance of δ and the covariance of δ and ϵ. Monte Carlo simulation studies show that analysis of data generated by the random coefficient model can give overestimated standard errors and higher mean squared error when applying regular regression with interactions, implying that important interactions may be overlooked.

The random coefficient model allows for a heteroscedastic residual variance. Whereas in regular regression the residual variance is assumed to be the same for all individuals, $V(y \mid x, z) = V(\epsilon)$, the residual variance for the random coefficient model varies with x. The conditional variance of y in (3) is

$$V(y_i \mid x_i, z_i) = V(\delta_i)\ x_i^2 + 2\ Cov(\delta_i, \epsilon_i)\ x_i + V(\epsilon_i). \tag{4}$$

This shows that the residual variance of (4) is a quadratic function of x. The two parameters specific to the random coefficient model, the variance of δ and the covariance of δ and ϵ, are estimated using information on the heteroscedasticity in the data corresponding to (4).

Random coefficient modeling can have more than one random slope. However, it should be mentioned that the random coefficient model does not always lead to an analysis that is as unproblematic as with the regular regression model. This is mainly due to fitting parameters to variances as indicated in (4). There has to be sufficient heteroscedasticity in the data. For example, the random coefficient model is not identified with a random slope for a binary covariate given that this provides only four pieces of information, the mean and variance for each covariate group, while the model has five parameters. Even with continuous

covariates a small degree of heteroscedasticity may lead to an almost flat log likelihood with random slopes so that convergence is difficult to achieve. The sparseness of information is clear when comparing to the use of random slopes in multilevel modeling. Unlike multilevel modeling, random coefficient modeling has only one observation available per individual, that is, one observation per cluster in multilevel terms. There are, however, designs where several observations are taken on y for the same value of x; see, e.g. Weisberg (2014, pp. 169-171).