Supplemental Material for Wiedermann, Artner, and von Eye, Heteroscedasticity as a Basis of Direction Dependence in Reversible Linear Regression Models, Multivariate Behavioral Research

Extensions to Categorical Covariates

To be placed on MBR website.

In this supplement, we discuss approaches to adjust directionally competing models for categorical covariates which are a priori known to be external influences. In general, we make use of the fact that any multiple linear regression model can be re-expressed as a (partial) regression model based on covariate-residualized variables (see, e.g., Karlson, Holm & Breen, 2012; Wurm & Fisicaro, 2014). We show that error terms of these two models are identical which implies that statistical inference on the error distribution of the partial regression model also holds for the multiple regression model and vice versa. Next, we show that non-independence of the error term and the predictors also holds for the directionally mis-specified partial regression model which enables researchers to deduce statements concerning the direction of effects of competing partial regression models. Results of a Monte-Carlo simulation study are presented which demonstrate the applicability of the proposed direction dependence methods to the case of categorical covariates.

For simplicity, we focus on the case of three variables: a true continuous outcome y, the true continuous (non-normally distributed) cause x, and a binary covariate z (with $z \in \{0,1\}$; the following proof can straightforwardly be extended to polytomous and ordinal predictors through defining proper sets of dummy variables). It is important to re-iterate that categorical variables are assumed to constitute *external influences* whose causes lie outside the considered model. That is, cases in which categorical variables serve as tentative outcomes of other explanatory variables are excluded. This imposes stricter assumptions on categorical covariates compared to the continuous case. However, the current set-up still allows model selection in multiple-group scenarios where the causal relation between predictor and outcome is allowed to vary in magnitude (i.e., in terms of the mean structure) across categorical groups which covers a broad

range of applications. In case of a linear relation between outcome and all explanatory variables, the true model is given by

$$y = i_y + b_{yx}x + b_{yz}z + \epsilon_y \tag{S1}$$

where i_y denotes the model intercept, the b's constitute the regression slopes, and ϵ_y refers to the error term which is assumed to be independent of x and z. In this case, the directionally mis-specified model can be written as

$$x = i_x + b_{xy}y + b_{xz}z + \epsilon_x. \tag{S2}$$

An alternative representation of the models in (S1) and (S2) can be obtained through residualizing x and y on z and regressing the corresponding residuals on each other. In other words, in a first step, one estimates the two auxiliary models

$$y = i'_y + b'_{yz}z + \epsilon'_y \tag{S3}$$

and

$$x = i'_x + b'_{xz}z + \epsilon'_x \tag{S4}$$

and extracts the residuals $\epsilon'_y = y - i'_y - b'_{yz}z$ and $\epsilon'_x = x - i'_x - b'_{xz}z$ which can be interpreted as "purified" measures of y and x (cf. Hayes, 2013). Next, the two competing models, $x \rightarrow y$ and $y \rightarrow x$, are obtained through

$$\epsilon'_y = a_{yx}\epsilon'_x + \theta_y \tag{S5}$$

and

$$\epsilon'_x = a_{xy}\epsilon'_y + \theta_x. \tag{S6}$$

It is well-known that the simple slope coefficients in (S5) and (S6) equal the partial regression coefficients given in (S1) and (S2), i.e., $b_{yx} = a_{yx}$ and $b_{xy} = a_{xy}$. Further, the two models that posit that $x \rightarrow y$, (S1) and (S5), have identical error terms ($\epsilon_y = \theta_y$) and so do the two competing models (S2) and (S6), i.e., $\epsilon_x = \theta_x$. Equivalence of errors can easily be shown through rewriting the error terms as function of observed variables. For model (S5), for example, one obtains

$$\begin{aligned} \theta_y &= \epsilon'_y - a_{yx}\epsilon'_x \\ &= y - \left(i'_y - b_{yx}i'_x\right) - b_{yx}x - (b'_{yz} - b_{yx}b'_{xz})z \end{aligned} \tag{S7}$$

with $i_y = i'_y - b_{yx}i'_x$ and $b_{yz} = b'_{yz} - b_{yx}b'_{xz}$ which is identical to ϵ_y in model (S1). Similar considerations hold for the two directionally competing models.

Further, because both true predictors, x and z, are assumed to be independent of ϵ_y, independence will also hold for the correctly specified partial regression model, i.e., ϵ'_x and θ_y (= ϵ_y) are stochastically independent in the present set-up. In contrast, non-independence of ϵ'_y and θ_x (= ϵ_x) can again be established through making use of the reverse corollary of the Darmois-Skitovich theorem: Here, ϵ'_y and θ_x can be expressed as linear functions of the same random variates (ϵ'_x and ϵ_y), i.e.,

$$\epsilon'_y = b_{yx}\epsilon'_x + \epsilon_y \tag{S8}$$

and

$$\begin{aligned} \theta_x &= \epsilon'_x - b_{xy}\epsilon'_y \\ &= \left(1 - b_{xy}b_{yx}\right)\epsilon'_x - b_{xy}\epsilon_y. \end{aligned} \tag{S9}$$

Because both, x and ϵ'_x, are assumed to be non-normally distributed, the "purified" outcome ϵ'_y and the error term of the mis-specified auxiliary regression model θ_x are stochastically non-independent when $\left(1 - b_{xy}b_{yx}\right)b_{yx} \neq 0$. From a direction dependence perspective, we, thus, arrive at the conclusion that a model of the form $x \rightarrow y$ is empirically confirmed, if independence holds in model (S5) and, at the same time, non-independence is observed in model (S6). In contrast, one has found empirical evidence for $y \rightarrow x$ if independence holds in model (S6) and, at the same time, non-independence can be observed in the model (S5). Again, if the independence assumption is satisfied/violated in both models, no distinct decision can be made.

Simulating the Performance of Heteroscedasticity Tests in Case of a Categorical Covariate

In this section, we present additional simulation results on the adequacy of the proposed direction dependence approach (in terms of Type I error and power) when categorical covariates are considered. In addition, we compare two different approaches to establish statements concerning the direction of effects: 1) the ordinary multiple linear regression approach (as used in case of continuous covariates) and 2) the partial regression approach where residuals of auxiliary regressions are used as covariate-adjusted ("purified") measures.

Type I Error Simulation

Data were generated according to the true model $y = i_y + b_{yx}x + b_{yz}z + \epsilon_y$ with x being a continuous variable and z denoting a binary covariate. The intercept was fixed at zero

and the error term ϵ_y was randomly sampled from the standard normal distribution. To generate correlated predictor variables, we, in the first step, generated two continuous variables, x and z' (exhibiting zero means and unit variances), using elliptical copulas of the normal family and, in a second step, dichotomized z'at the theoretical mean of zero, i.e., z = 0 if $z' \leq 0$ and z = 1 otherwise. Because dichotomization affects the correlation between x and z, input correlations between x and z' were adjusted using $\rho_{xz} = \rho_{xz'}(h/\sqrt{p(1-p)})$ where h represents the ordinate of the standard normal curve at the point of dichotomization and p denotes the proportion of observations for which $z' > 0$ (cf. Cohen, 1983; MacCallum et al., 2002). We restricted the simulation to the case of equal group sizes (i.e., p = 0.5 and h = .3989). Values for $\rho_{xz'}$ were selected to obtain the desired correlations of ρ_{xz} = 0, 0.2, 0.4, and 0.6. Regression coefficients were selected to account for zero, small, medium, and large effects reflecting partial correlations of 0, 0.14, 0.36, and 0.51. For the continuous predictor x, we used b_{yx} = 0, 0.14, 0.39, and 0.59; corresponding values for the binary covariate z were b_{yz} = 0, 0.28, 0.77, and 1.19. Sample sizes were *n* = 50, 100, 200, and 400. The simulation factors were fully crossed leading to 4 (magnitude of ρ_{xz}) × 4 (magnitude of b_{yx}) × 4 (magnitude of b_{yz}) × 4 (sample size *n*) = 256 experimental conditions (1000 samples were generated per condition).

For each variable triple (x, y, and z), two different approaches were used: First, competing multiple regression models $y = i_y + b_{yx}x + b_{yz}z + \epsilon_y$ and $x = i_x + b_{xy}y + b_{xz}z + \epsilon_x$ were estimated and the Breusch-Pagan test was used to evaluate the homoscedasticity of ϵ_y and ϵ_x using a nominal significance level of 5% (we focus on the Breusch-Pagan test which has been shown to be the most powerful procedure to establish direction dependence decisions). The empirical Type I error rates were defined as the portion of samples for which $H_0: \Omega_{\{x,z\}\to y} = I_n$ was retained and, at the same time, $H_0: \Omega_{\{y,z\}\to x} = I_n$ was rejected.

Second, we estimated the two auxiliary regression models $y = i'_y + b'_{yz}z + \epsilon'_y$ and $x = i'_x + b'_{xz}z + \epsilon'_x$ and corresponding residuals (ϵ'_y and ϵ'_x) were further analyzed using the two simple linear regression models $\epsilon'_y = a_{yx}\epsilon'_x + \theta_y$ (reflecting $x \to y$) and $\epsilon'_x = a_{xy}\epsilon'_y + \theta_x$ (reflecting $y \to x$). Again, the Breusch-Pagan test was used to evaluate the homoscedasticity assumption in both models separately. Here, the empirical Type I error rate is defined as the

portion of samples for which $H_0: \Omega_{\{\epsilon'_x\}\to\epsilon'_y} = I_n$ is retained and, at the same time, $H_0: \Omega_{\{\epsilon'_y\}\to\epsilon'_y} = I_n$ is rejected.

Figure S1 summarizes the empirical Type I error rates of both regression approaches as a function of b_{yx}, b_{yz}, and ρ_{xz} for the case of n = 50 observations. Results for larger sample sizes are virtually identical and will, thus, not be presented here. Type I error rates are given for the separate Breusch-Pagan tests and the combined decision in terms of model selection for the multiple regression approach and the two-step auxiliary regression approach. The dashed horizontal lines reflect Bradley's (1978) liberal robustness interval. In general, observed Type I error rates for both approaches are in accordance with the nominal significance level of 5%.

Power Simulation

Next, we focused on the power behavior of the proposed approach when adjusting for a categorical covariate. The true (continuous) predictor, x, was generated from various gamma distributions (shape parameter $c = (2/\gamma)^2$, scale parameter d = 1) with pre-specified skewness values of γ_x = 0.75, 1.5, and 2.25. Again, the error term of the true model (ϵ_y) was sampled from the standard normal distribution and the binary covariate (z) was generated through dichotomizing a standard normal variable (z') at the mean of zero. The simulation experiment consisted of 4 (magnitude of ρ_{xz}) × 4 (magnitude of b_{yx}) × 4 (magnitude of b_{yz}) × 3 (skewness of x) × 4 (sample size n) = 768 conditions. Again, 1000 samples were generated for each condition. Multiple linear regression ($\{x, z\} \to y$ versus $\{y, z\} \to x$) as well as the two-step auxiliary regression approach ($\epsilon'_x \to \epsilon'_y$ versus $\epsilon'_y \to \epsilon'_x$) were applied for model selection purposes. For each simulation condition, we calculated the portion of correct model selection decisions, i.e., the portion of cases in which the Breusch-Pagan test indicated homoscedasticity for the correctly specified model and heteroscedasticity for the mis-specified model (for the Breusch-Pagan tests we again used a nominal significance level of 5%). That is, the multiple regression approach identifies the correct causal flow when $H_0: \Omega_{\{x,z\}\to y} = I_n$ is retained and, at the same time, $H_0: \Omega_{\{y,z\}\to x} = I_n$ is rejected. Similarly, the auxiliary regression approach selected the correct model when $H_0: \Omega_{\{\epsilon'_x\}\to\epsilon'_y} = I_n$ is retained and, at the same time, $H_0: \Omega_{\{\epsilon'_y\}\to\epsilon'_y} = I_n$ is rejected.

Figure S2 shows the empirical rejection rates of separate Breusch-Pagan tests for the multiple and auxiliary regression models as well as the empirical power curves in terms of model selection based on combined decisions. Note that empirical rates of individual Breusch-Pagan test evaluating $H_0: \Omega_{\{x,z\}\rightarrow y} = I_n$ and $H_0: \Omega_{\{\epsilon'_x\}\rightarrow \epsilon'_y} = I_n$ again reflect Type I error rates because data were simulated such that the homoscedasticity assumption holds in the correctly specified models $\{x, z\} \rightarrow y$ and $\epsilon'_x \rightarrow \epsilon'_y$. The remaining curves in Figure S2 reflect observed power values for rejecting homoscedasticity in the mis-specified model and for combined model selection decisions for the multiple and auxiliary regression models. Overall, model selection results are virtually identical compared to the case of a continuous covariate (see Figure 3 of the main text). In other words, the power to select the correct model for both, the multiple and the auxiliary regression approach, increases with the magnitude of b_{yx}, the skewness γ_x, and sample size n. Further, the power also slightly increases with the correlation between x and z. In contrast, the magnitude of b_{yz} has no impact on the model selection procedures. In general, the multiple linear regression approach has a slight power advantage over the auxiliary regression approach.

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253. doi: 10.1177/014662168300700301

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford.

Karlson, K. B., Holm, A., & Breen, R. (2012). Comparing regression coefficients between same-sample nested models using logit and probit. A new method. *Sociological Methodology, 42*, 286-313. doi: 10.1177/0081175012444861

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19-40. doi: 10.1037/1082-989x.7.1.19

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72*, 37-48. doi: 10.1016/j.jml.2013.12.003

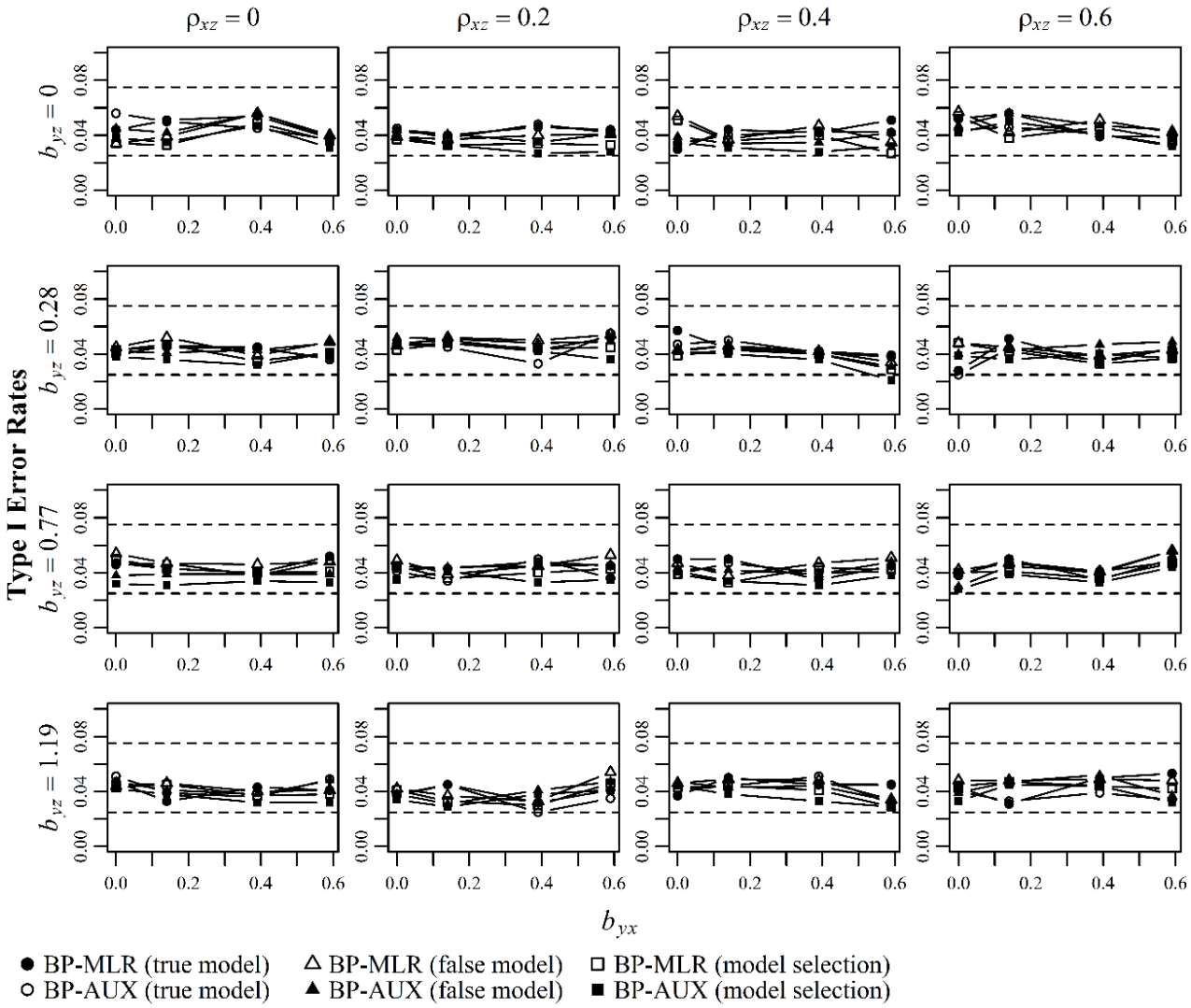

Figure S1: Empirical Type I error rates of the Breusch-Pagan test (BP) for multiple linear regression (MLR) and auxiliary regression models (AUX). Type I error curves labeled with “true model“ refer to models with correctly specified direction of effects, curves labeled with “false model“ refer to directionally mis-specified models. “model selection“ refers to the Type I error rates of combined homoscedasticity decisions.

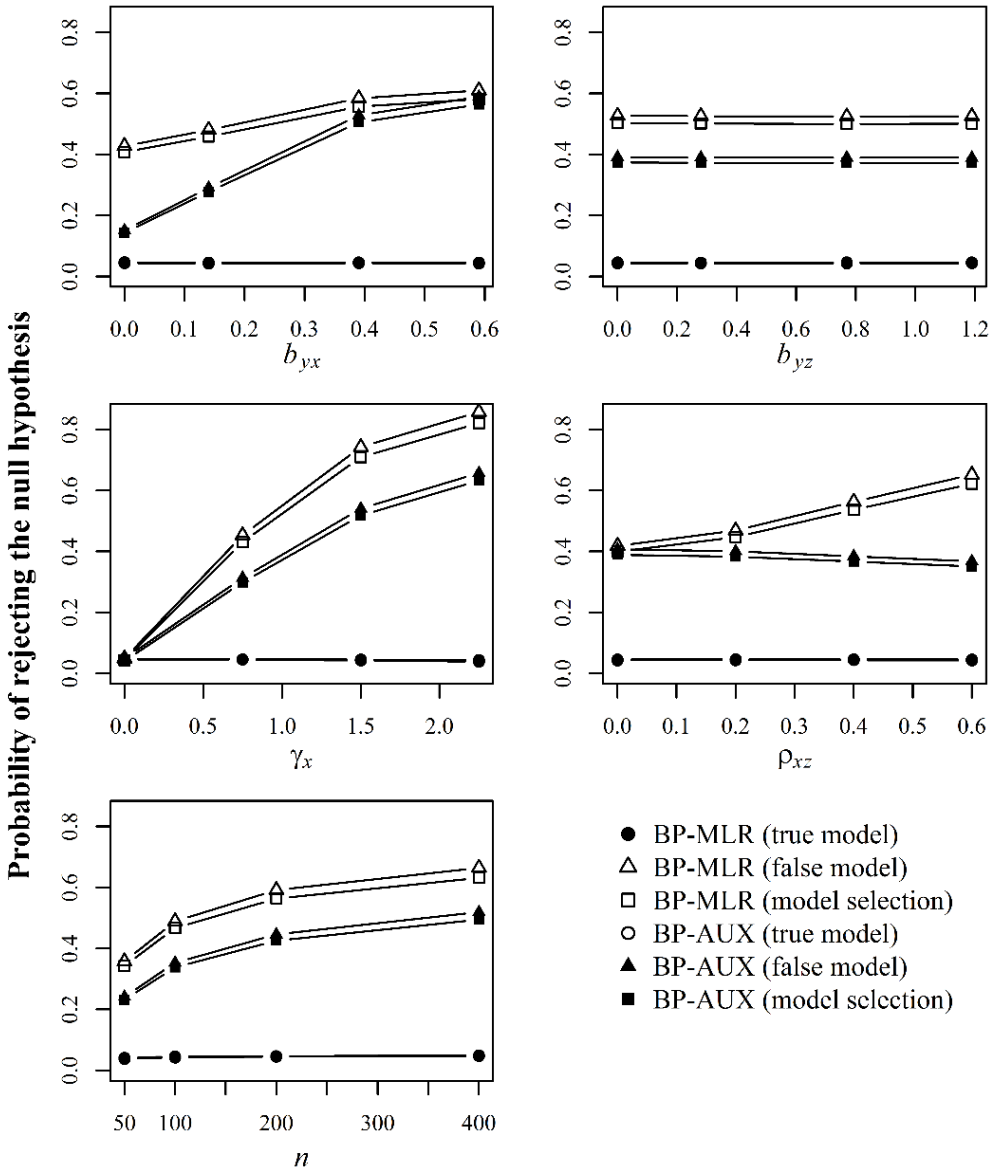

Figure S2: Empirical rejection rates of the Breusch-Pagan test (BP) for multiple linear regression (MLR) and auxiliary regression models (AUX). Curves labeled with "true model" refer to individual models with correctly specified direction of effects and reflect Type I error rates of the Breusch-Pagan test, curves labeled with "false model" refer to directionally mis-specified models and reflect the power of the Breusch-Pagan test to identify heteroscedasticity. Curves labeled with "model selection" refer to the power of the model selection procedure based on combined homoscedasticity decisions.