Chapter 2

Daniel C. Furr

January 9, 2016

Contents

1 Introduction

2 Simulation and analysis methods

2.1 Data generation

Data are simulated for varying numbers of persons (P) and items (I) using the model described in Chapter 1. Specifically, the composite item difficulties are generated as

$$\delta_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{2i}x_{3i}\beta_5 + \epsilon_i \tag{1}$$

$$\epsilon_i \sim \mathrm{N}(0, \tau^2), \tag{2}$$

where $x_{1i} = 1$ is an intercept and x_{2i}, x_{3i}, and x_{4i} are indicator variables which each equal 1 for half of the items and 0 for the remainder. Each possible combination of the indicators occurs an equal number of times, and the generating model includes one interaction, $x_{2i}x_{3i}$. Table 1 provides the design matrix for item covariates. The rows of the design matrix are repeated to accommodate multiples of 8.

Table 1: Items design matrix

x_1	x_2	x_3	x_4
1	0	0	0
1	0	0	1
1	0	1	0
1	0	1	1
1	1	0	0
1	1	0	1
1	1	1	0
1	1	1	1

The composite abilities are generated as

$$\theta_p = w_{1p}\gamma_1 + w_{2p}\gamma_2 + \zeta_p \tag{3}$$

$$\zeta_p \sim \mathrm{N}(0, \sigma^2), \tag{4}$$

where w_{1p} and w_{2p} are also crossed indicator variables. Table 2 presents the design matrix for person covariates, the rows of which are repeated to accommodate the $P = 4n$ persons.

A key feature of the generated datasets (and data of this type more generally) is the extent to which the item covariates account for the composite item difficulties. To this end, let $\upsilon^2 = \mathrm{var}(x'\beta)$ represent the variance of the structural part of item difficulty. Because of the item design, υ^2 does not vary between simulated datasets, even if they have differing numbers of items (so long as I is a multiple of 8.). The total item variance is $\upsilon^2 + \tau^2$. Then

$$R^2 = \frac{\upsilon^2}{\upsilon^2 + \tau^2} \tag{5}$$

Table 2: Persons design matrix

w_1	w_2
0	0
0	1
1	0
1	1

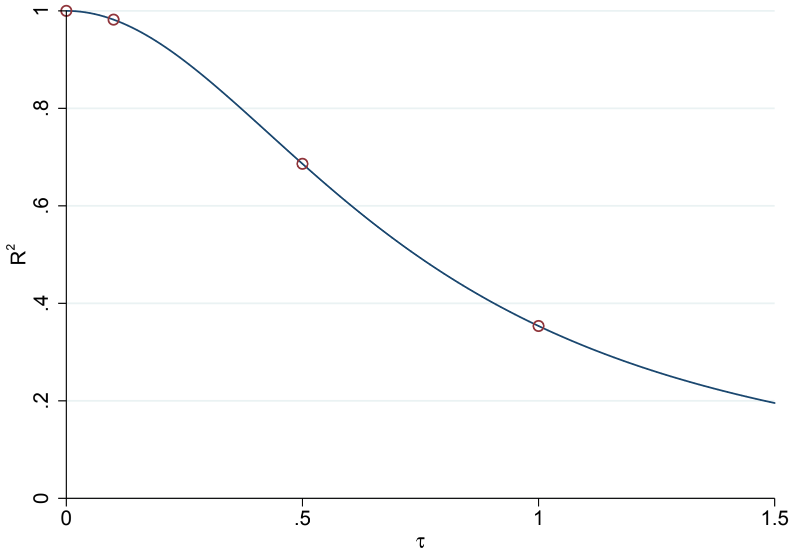

Figure 1: R^2 versus τ for the simulations. The points indicate generating values of τ.

represents the proportion of item variance accounted for by the item predictors.

The generating values for the structural part of item difficulties are $\beta = \{-.5, .5, .5, .5, -.5\}$ in all simulation conditions, and so $\upsilon^2 = 0.55$ in all conditions. Figure 1 displays R^2 as a function of τ with υ^2 fixed to this value. The points marked indicate the generating values of τ, which are $\tau \in \{0.00, 0.10, 0.50, 1.00\}$ (or equivalently, $\tau^2 \in \{0.00, 0.01, 0.25, 1.00\}$). This choice yields $R^2 \in \{1.00, 0.98, 0.69, 0.35\}$. On the person side, the generating values are fixed across conditions, with $\gamma = \{.5, .5\}$ and $\sigma = 1$. The numbers of persons and items vary: $I \in \{32, 128\}$ and $P \in \{300, 1000\}$.

In summary, three factors are varied between simulation conditions in a crossed design: τ (and by extention, R^2), I, and P. All other elements are fixed across conditions. Because cross-validation features prominently in this chapter, multiple datasets are simulated within each replication. A “training” dataset is created as described above, and along with it three “test” datasets are formed: one representing a sample with new items (corresponding to new draws of ϵ_i), one representing a sample with new persons (new draws of ζ_p), and the last representing a sample with both new items and new persons.

2.2 Models

Three models, differing only in specification of δ_i, are fit. Model 1 includes only the "main effects" for the item covariates:

$$\delta_i^{(1)} = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4. \tag{6}$$

Model 2 adds an interaction:

$$\delta_i^{(2)} = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{2i}x_{3i}\beta_5. \tag{7}$$

Model 3 adds an additional, extraneous interaction:

$$\delta_i^{(3)} = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{2i}x_{3i}\beta_5 + x_{3i}x_{4i}\beta_6. \tag{8}$$

None of the analysis models includes the residual ϵ_i. Each analysis model models ability as in Equations 3 and 4.

3 Naive cross-validation methods

One naive approach to model selection is the use of significance testing for parameters. In order to select among the three analysis models, a researcher may fit Model 2 and make a judgment based on the p-value for β_5, the parameter associated with the interaction. If non-significant, the researcher may select Model 1. Otherwise, the researcher may fit Model 3. If the additional interaction (β_6) is significant, Model 3 would be selected. Otherwise, Model 2 would be selected. This is a forward stepwise procedure.

This method works well when τ is small but poorly otherwise. When $\tau = 0$, Model 2 matches the data generating model exactly, and Models 1 and 3 are close. The result is that the correlated nature of responses within an item cluster are appropriately accounted for, yielding correct standard errors for β. The preceding is approximately true for small values of τ like $\tau = .1$. However, with greater values of τ, the analysis models fail to account for the within-cluster dependency, resulting in standard errors that are too low. This shortcoming leads to Model 3 being selected the majority of the time when τ takes medium to large values. [*Add back in results supporting this or remove.*]

Adding item residuals ϵ_i to the analysis models would provide correct standard errors and p-values. Such a model is prohibitively difficult to fit without resorting to Monte Carlo methods, though. Further, in practical application there may be many more than three models under consideration, which brings up complexities around multiple hypothesis testing. For this reason an appealing alternative is AIC, defined as

$$\text{AIC} = \text{deviance} + 2k, \tag{9}$$

where k is the number of model parameters. The model with the lowest value of AIC is selected. The results of using AIC with the simulated datasets are presented in Figure 2.

AIC is an approximation for holdout cross-validation, in which a model is estimated using a "training" dataset and then evaluated on a "test" dataset. In this instance, AIC

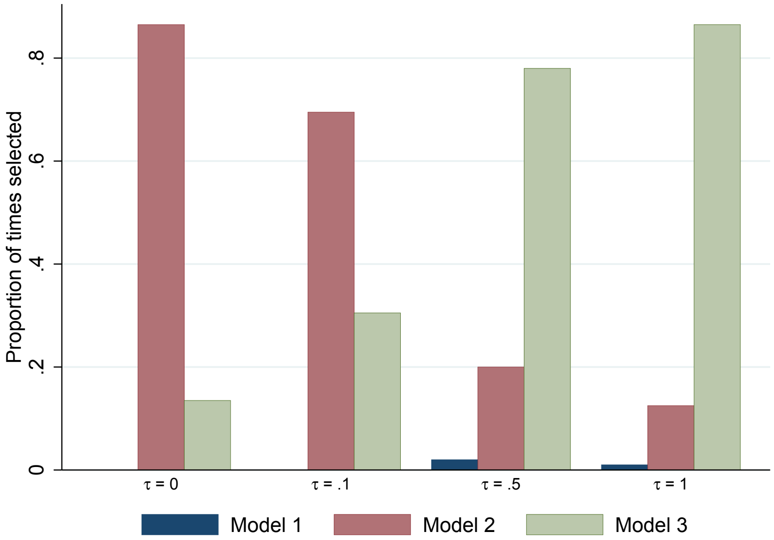

Figure 2: Proportion of times each model was selected using AIC.

approximates the deviance that would result from applying the trained model to a test dataset consisting of new persons and the same items. [*Explain why, from chapter 1 or cite papers.*] k in Equation 9 may be viewed as an adjustment to the deviance of the model fit to the training data owing to uncertainty in the parameter estimates.

The correct values for k may be estimated from the simulation by subtracting the deviance from the model fit to the training data from the fit to test data consisting of new persons responding to the same items. This is presented in Figure 3. The empirical estimates for k are similar to those given by AIC (7.00, 8.00, and 9.00) with some variation across values for τ. Importantly, the difference between models is consistent across values for τ and with AIC.

In short, AIC performs poorly for model selection in this instance even though it accurately estimates the out-of-sample deviance. Further, cross-validation with new persons also performs poorly, as depicted in Figure 4. Lastly, BIC performs somewhat better but has the same problem, as shown in Figure 5.

4 Holdout cross-validation for item predictors

If the focus of model selection is the choice of item predictors, cross-validation schemes based on test data with the same items are wrongheaded. A useful approach instead is to consider how the item predictors will fare for a new set of items constructed from the same item design. To this end, the analysis models are fit to a training dataset and then evaluated on a holdout dataset representing the same persons and new items. Figure 6 provides the proportion of times each model was selected using this scheme.

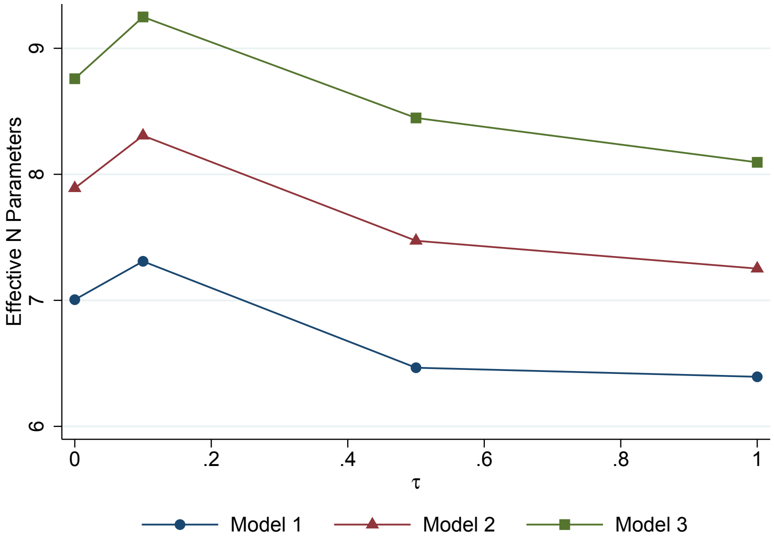

Figure 3: Estimated effective number of parameters for models fit to test data consisting of new persons responding to the same items.

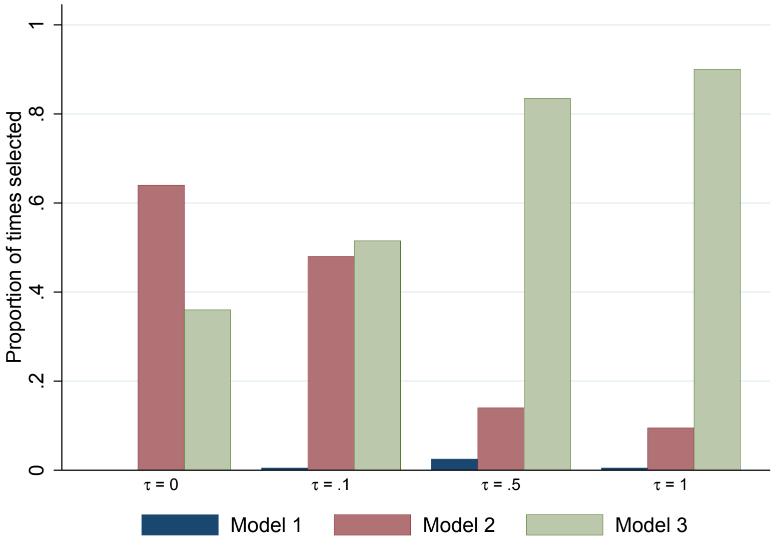

Figure 4: Proportion of times each model was selected using cross-validation with new perons.

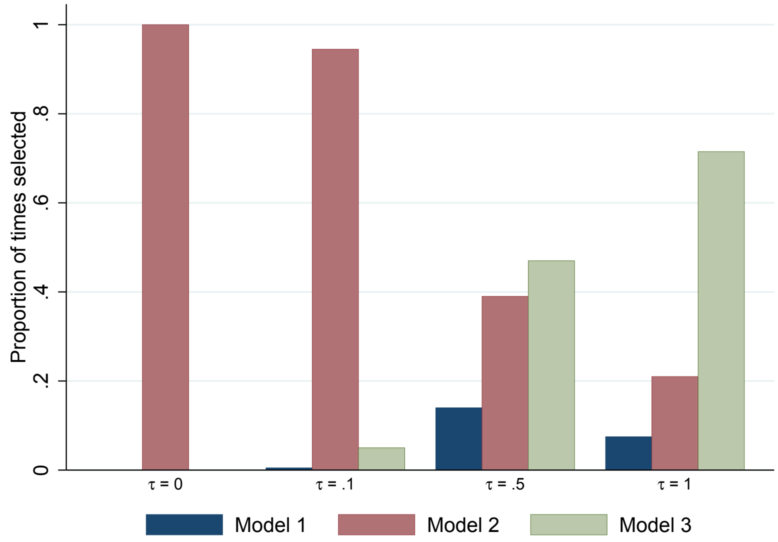

Figure 5: Proportion of times each model was selected using BIC.

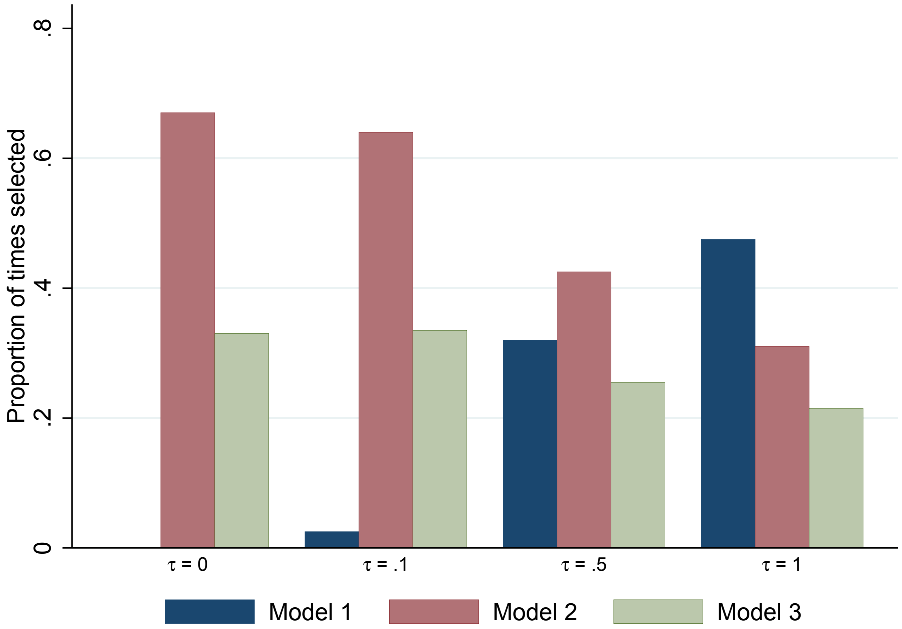

Figure 6: Proportion of times each model was selected using holdout cross-validation with holdout data consisting of the same persons and new items.

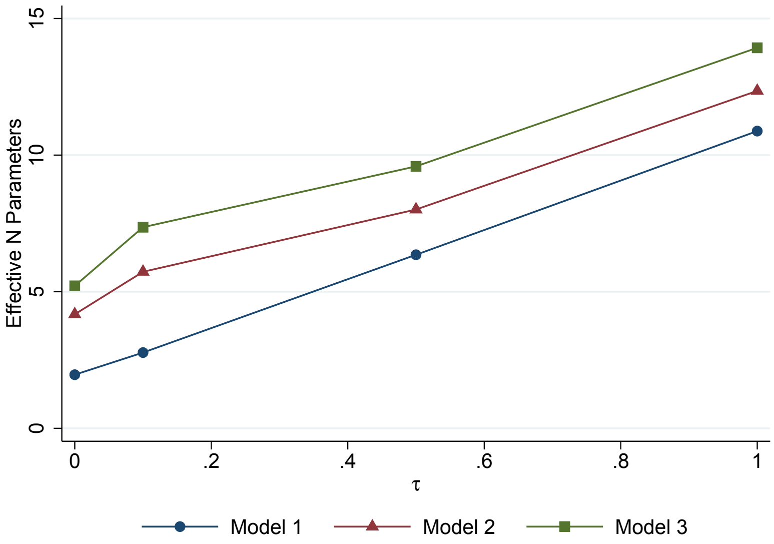

Figure 7: Estimated effective number of parameters for models fit to test data consisting of the same persons responding to new items.

5 Cross-validation with a fixed-persons random-items model.

A variation on the three models is considered in which the persons are modeled as fixed effects and the items as random effects. The fixed part of the models include the item predictors and and indicator variable for each sum score. Person covariates are omitted. In these "inverted" models, the items are regarded as exchangeable rather than the persons. It is posited that AIC for the inverted models would perform correctly for cross-validation inferences that involve new items.

Figure 7 provides the estimated effective number of parameters for the models based on the simulations. As may be seen, the values are low in comparison with the count of parameters (35.00, 36.00, and 37.00) and dependent on τ. Figure 8 shows the proportion of times each inverted model was selected using AIC.

6 Discussion

[*Find papers on choosing between LLTMS.*]

[*(1) K-fold CV. (2) Consider a linear model to parallel holdout CV with new items. Possible to get marginal likelihood with linear model? (3) Show that AIC works with "De Boeck" version of model. (4) Consider extending topic to include linear crossed mixed-effects models.*]

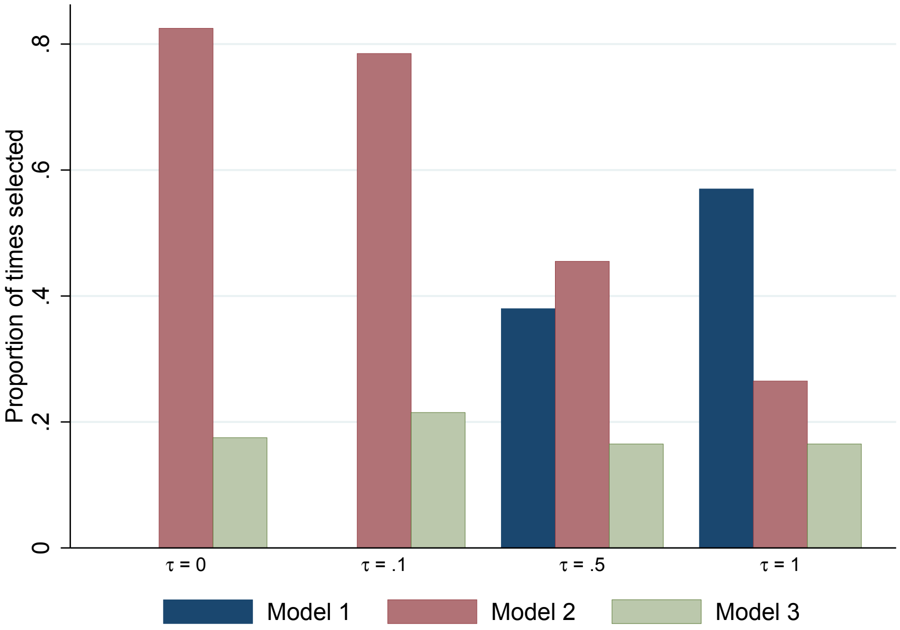

Figure 8: Proportion of times each model was selected using AIC with the “inverted” models.