Chapter 2

Daniel C. Furr

January 25, 2016

Contents

1 Introduction **2**

2 Simulation and analysis methods **2**
2.1 Data generation 2
2.2 Models 4

3 Simulation 1 **4**
3.1 Naive cross-validation methods 5
3.2 Holdout cross-validation for item predictors 7

4 Simulation 2 **7**

5 Discussion **8**

1 Introduction

The use of cross-validation with clustered data requires a decision as to what aspects of the data are exchangeable, and so are resampled in new data, and what aspects persist. For item response data, in which responses are nested both within persons and items simultaneously, new data may consist of a new group of persons responding to the original items, which will be referred to as cross-validation "over persons." Alternatively, new item response data may entail a new set of items presented to the original group of persons, which will be referred to as cross-validation "over items." Other variations are possible, such as new data from both the same persons and items, an exact replication of the original data collection, or new data from both new persons and items.

The form of new data determines what inferences may be made using cross-validation. For example, cross-validation must be done over persons in order to select for person ability predictors in a latent regression Rasch model (). In other words, a new sample of persons are needed to assess the predictors for ability. Likewise, if the goal is to choose predictors for item difficulty in a linear logistic test model (LLTM) (), cross-validation must be performed over items.

The choice of model has strong implications regarding which aspects of the data are considered exchangeable, and hence would be resampled in cross-validation, and which are persistent. Specifically, the "random effects" are exchangeable, and the "fixed effects" are persistent. In general for item response models, person abilities (or their residuals) are modeled as random effects, while items are treated as fixed, implying that a repeat of the data collection would involve a new sample of persons responding to the same items.

In this chapter, the difference in efficacy of employing cross-validation over persons versus over items with LLTM is investigated. AIC, a single dataset approximation to cross-validation, is shown to correspond to cross-validation over persons, which is not effective for the selection of item predictors.

2 Simulation and analysis methods

2.1 Data generation

Data are simulated 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 $I = 8n_p$.

The composite abilities are generated as

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

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

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

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 $P = 4n_i$ persons.

Table 2: Persons design matrix

w_1	w_2
0	0
0	1
1	0
1	1

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 (in multiples of eight). The total item variance is $\upsilon^2 + \tau^2$. Then

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

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 $\upsilon^2 = 0.11$. Figure 1 displays R^2 as a function of τ with υ^2 fixed to this value. The points indicate the generating values of τ, which are $\tau \in \{0.00, 0.10, 0.30, 0.50\}$ (or equivalently, $\tau^2 \in \{0.00, 0.01, 0.09, 0.25\}$). These values yield $R^2 \in \{0.30, 0.55, 0.92, 1.00\}$. On the person side, the generating values are fixed across conditions, with $\gamma = \{.5, .5\}$ and $\sigma = 1$.

In summary, the same generating values and design matrices are used across all simulation conditions. In the first simulation, generating values of τ are varied. In the second, the number of items are varied. To employ cross-validation, multiple datasets are simulated within each replication: the "test" dataset, a "training" dataset representing a sample with new items (corresponding to new draws of ϵ_i), and another training dataset representing a sample with new persons (new draws of ζ_p).

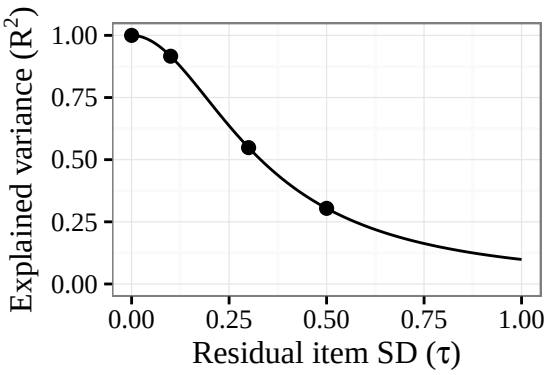

Figure 1: Proportion of total item variance explained by the item covariates (R^2) versus residual item standard deviation (τ) in the simulated datasets. The points indicate generating values of $\tau \in \{0.00, 0.10, 0.30, 0.50\}$.

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:

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

Each analysis model models ability as in Equations 3 and 4 while omitting the item residual ϵ_i. Model 2 has the correct fixed part of the model, and consequently Model 2 matches the data generating model when $\tau = 0$. Otherwise, none of three match the data generating model.

The analysis models are naive, given that in most simulation conditions $\tau > 0$. More generally, the item predictors in an LLTM are not expected to exactly fit the item difficulties in actual application. Omitting item residuals is overly restrictive and results in failure to model the within-item dependency of responses. This problem is the same as that which arises from fitting a non-hierarchical model to clustered data. However, fitting random effects for both person and items is prohibitively difficult with maximum likelihood estimation, so researchers often resort to fitting models like the above.

3 Simulation 1

Datasets consisting of 500 persons and 32 items are generated with values of $\tau \in 0.00, 0.10, 0.30, 0.50$. The simulation is carried out for 500 replications for each value of τ.

3.1 Naive cross-validation methods

One naive approach to model selection is the use of significance testing for parameters. In order to select among nested analysis models, a researcher may conduct likelihood ratio tests on pairs of models. If the test rejects the simpler model in a pair, the more complex model is retained and compared against a still more complex model. This procedure continues until the test fails to reject a model or when only the most complex model remains.

The first panel of Figure 2 provides the proportion of times each model is selected using likelihood ratio tests in the simulation. Model 2 is selected the large majority of times when τ is zero or small, corresponding to situations in which this model is the true model or almost so. As τ becomes greater, the more complex Model 3 is selected increasingly often. The same results would be expected if Wald tests on the interaction terms were used for selection instead of likelihood ratio tests.

An appealing alternative is AIC, defined as

$$\text{AIC} = \text{dev}_{\text{in}} + 2k_{\text{AIC}}, \tag{9}$$

where k_{AIC} is the number of model parameters. The model with the lowest value of AIC is selected. AIC is suitable both for comparing non-nested models, which likelihood ratio tests cannot do, and for comparing many models simultaneously, which bring up problems of multiple hypothesis for likelihood ratio tests. Further, AIC explicitly frames model selection as a trade off between model fit and complexity. [*Explain further, including Kullback-Leibler divergence.*] 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 [*show Kullback-Leibler divergence formula*]. Specifically,

$$k_{\text{AIC}} \approx k_{\text{CV}} = \frac{\text{dev}_{\text{out}} - \text{dev}_{\text{in}}}{2}, \tag{10}$$

where dev_{in} is the deviance of the fitted model in the training data, and dev_{out} is the deviance of the model in the test data given parameter estimates obtained from the training data. [*Equation is vague. Is it for one application of CV, or a long run quantity.*] For the LLTM, k_{AIC} approximates the k_{CV} that results from test data consisting of a new sample of persons responding to the same items, or in other words, cross-validation over persons. [*Explain why, from chapter 1 or cite papers.*]

The correct values for k_{CV} from cross-validation over persons may be estimated from the simulation. This is presented in the left panl of Figure 3. The empirical results regarding k_{CV} are similar to those given by AIC (7, 8, and 9) with some variation across values for τ. Importantly, the difference between models is consistently about one, which agrees 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 3. Lastly, BIC performs somewhat better but has the same problem. BIC is associated with k_{BIC} being 33.88, 38.72, and 43.56 for the three models. [*Discuss BIC.*]

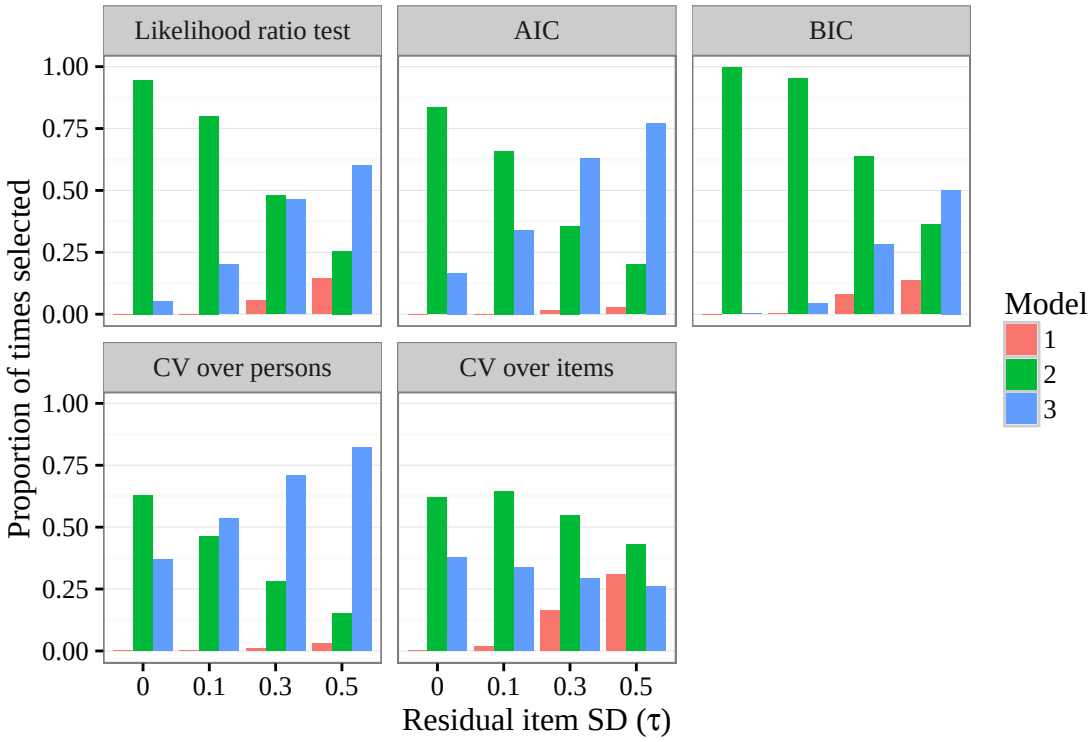

Figure 2: Proportion of times each model was selected across differing generating values for residual item standard deviation (τ), shown for differing selection methods. Each simulation condition was replicated 500 times with datasets consisting of 500 persons and 32 items. Model 2 is the true model when $\tau = 0$.

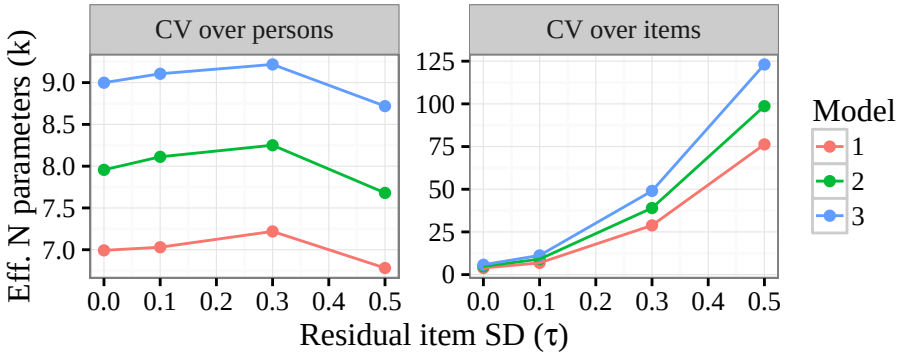

Figure 3: Estimated effective number of parameters (k) for the models by value of τ, shown for two cross-validation methods. k is the difference between the out-of-sample and in-sample deviance divided by two. AIC relies on an approximation of this quantity, in this case 7, 8, and 9, respectively. Each simulation condition was replicated 500 times with datasets consisting of 500 persons and 32 items. Model 2 is the true model when $\tau = 0$.

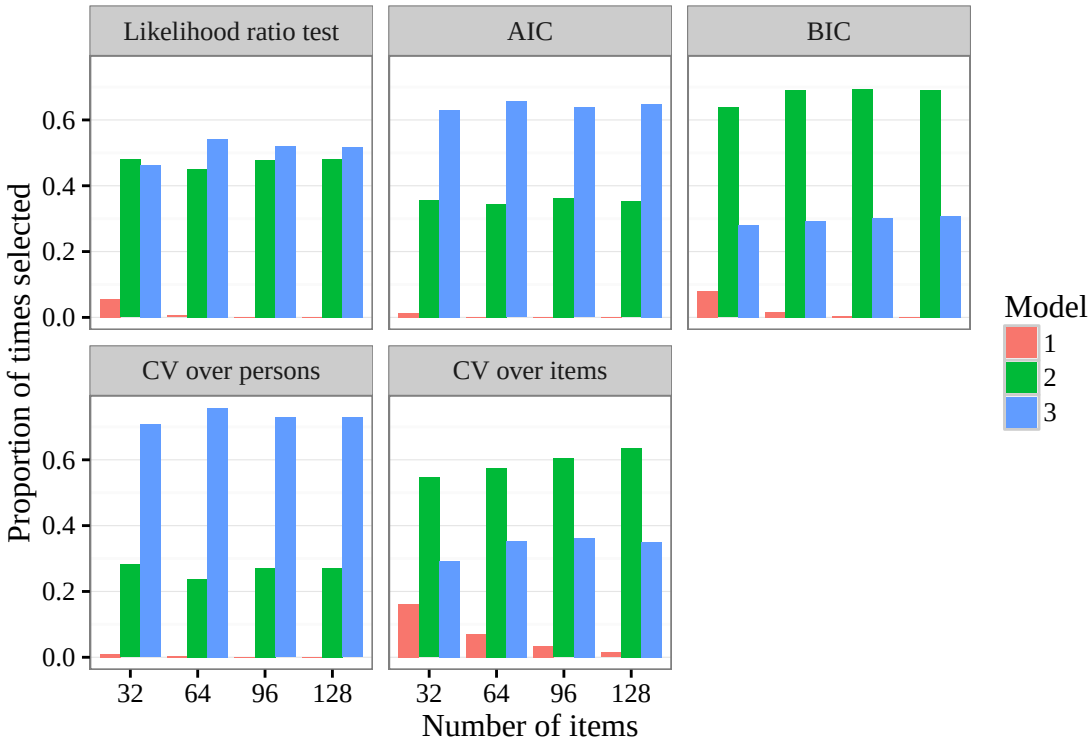

Figure 4: Proportion of times each model was selected depending on the number of items, shown for differing selection methods. Each simulation condition was replicated 500 times with datasets consisting of 500 persons and with $\tau = .3$.

3.2 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 with the same item covariates. 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 2 provides the proportion of times each model was selected using this scheme. Figure 3 provides k_{CV} under this scheme.

4 Simulation 2

The effect of differing numbers of items are considered. Datasets consisting of $P = 500$ persons and $I \in \{32, 64, 96, 128\}$ items are generated with values of τ fixed at 0.30. The simulation is carried out for 500 replications for each value of I. Figure 4 shows the proportion of times each model is selected depending on the number of items. Figure 5 provides empirical estimates of the effective number of parameters.

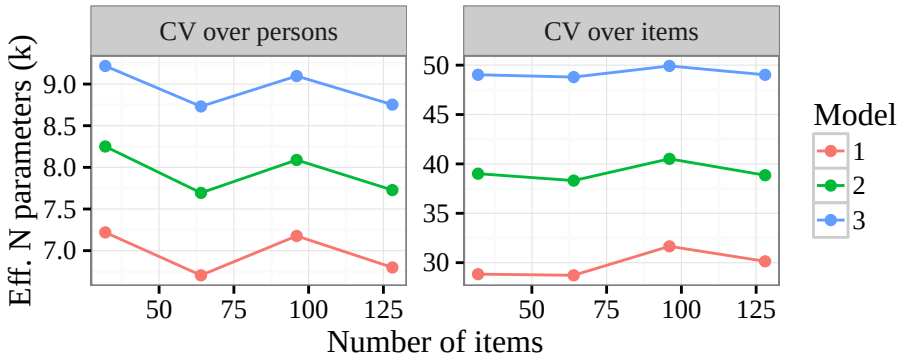

Figure 5: Estimated effective number of parameters (k) for the models by number of items, shown for two cross-validation methods. k is the difference between the out-of-sample and in-sample deviance divided by two. AIC relies on an approximation of this quantity, in this case 7, 8, and 9, respectively. Each simulation condition was replicated N times with datasets consisting of N persons and with $\tau = .3$

5 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.*]