Chapter 12
General and Specific Factors in Selection Modeling

Bengt Muthén

Abstract This chapter shows how analysis of data on selective subgroups can be used to draw inference to the full, unselected group. This uses Pearson-Lawley selection formulas which apply to not only regression analysis but also structural equation modeling. The chapter shows the connection with maximum-likelihood estimation with missing data assuming MAR versus using listwise deletion. Applications are discussed of selection into the military using factor analysis models for the variables used in the selection.

12.1 Introduction

Modeling with selective subgroups needs adjustments to be able to draw inference to the full group. This is a typical feature in predictive validity studies where a criterion outcome is regressed on or correlated with a predictor variable and the criterion outcome is missing for those not selected. The adjustments draw on Pearson-Lawley selection formulas (Pearson 1903; Lawley 1943–1944; Lord and Novick 1968; Johnson and Kotz 1972) to obtain desired inferences. The Pearson-Lawley formulas assume linear, homoscedastic regression of a set of analysis variables on a set of selection variables. The general Pearson-Lawley selection formulas can be used for deriving means, variances, and covariances for the full population given values of the selected population and vice versa. This chapter shows that Pearson-Lawley selection formulas play a role not only with respect to predictive validity assessment, but also with respect to multiple-group latent variable modeling. The connection between selection and maximum-likelihood estimation under the MAR assumption is illustrated by Monte Carlo simulations and real-data analyses.

B. Muthén (✉)
Graduate School of Education and Information Studies, Social Research Methodology Division, University of California, Los Angeles, USA
e-mail: bmuthen@ucla.edu

M. Rosén et al. (eds.), *Cognitive Abilities and Educational Outcomes*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-43473-5_12

Individuals applying for a certain training program may be selected based on a set of tests and other assessments. For example, students are selected into colleges based on the SAT, GRE, GMAC, or GMAT and job candidates are selected based on personality tests. To understand the quality of such a selection procedure, the tests and assessments are used as predictors of a training program outcome such as grades or job performance. The multiple correlation R value from this regression is viewed as a predictive validity coefficient. The estimation of this coefficient requires data on the program outcome and the predictors, which are available only for those who were selected. The interest is, however, in estimating the coefficient for the population of all applicants, not only those who were selected. Those who were selected are not a random subsample of those who applied, which means that the inference is distorted unless corrections are made. Similarly, screening instruments are used at baseline in psychological studies to determine a subsample that is at risk for certain future behavioral problems and is therefore of interest to follow up for further study. Again, the desired inference is to the population from which the baseline sample is taken, not to the subpopulation that is at risk.

12.2 Predictive Validity in a Simple Example

Consider the following linear regression

$$y_i = \alpha + \beta x_i + \varepsilon_i. \tag{12.1}$$

In a predictive validity context the predictor x is a test score used to select individuals into a training program in which a criterion outcome y is measured at the end of training. Selecting on x, the regression of y on x obtained in the group of selected individuals correctly estimates the regression model for the full, applicant population (see, e.g. Muthén and Joreskog 1983; Dunbar and Linn 1991). In contrast, selecting on y results in biased regression estimates. The two selection cases are illustrated in Fig. 12.1 using the example of Dunbar and Linn (1991) with a regression of y on x using standardized variables with correlation 0.6. The regression model was estimated using (a) a full sample of 5000 subjects, (b) a selected subsample of subjects with x scores above the mean, and (c) a selected subsample of subjects with y scores above the mean.

12.3 Monte Carlo Study of Selection in an SEM

In structural equation modeling, analysis of a selective group typically gives distorted estimates of the parameters for the full group. It is instructive to study the magnitude of such distortions through an example.

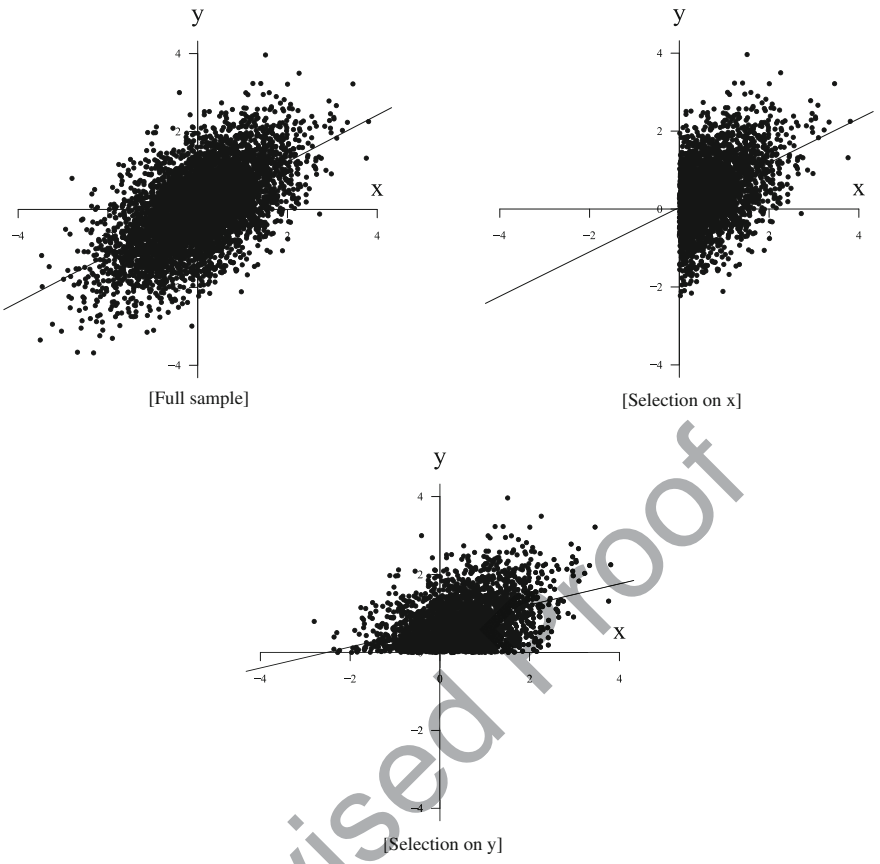

Fig. 12.1 Regression analysis using three different samples

Consider a latent variable version of the selection case of (2). Figure 12.2 corresponds to a hypothetical situation of a selection or screening measurement instrument formed by y_1–y_6, which are indicators of a general factor g and a specific factor s in line with bi-factor modeling. At a later time point a criterion measurement instrument y_7–y_{10} measures a single factor f_2. Consider first the case where the selection variable consists of the unweighted sum of y_1–y_6 so that those with the highest sum form the selected group which are followed up and administered the criterion test. Figure 12.3 shows the data structure, where the unselected group do not have observations on y_7–y_{10}.

The effects of selection on the analysis are illustrated by the following Monte Carlo simulation. A random sample of 2000 subjects is given the y_1–y_6 test and those with the top 50 % summed score are selected and given the y_7–y_{10} test. This procedure is repeated over 500 Monte Carlo replications.

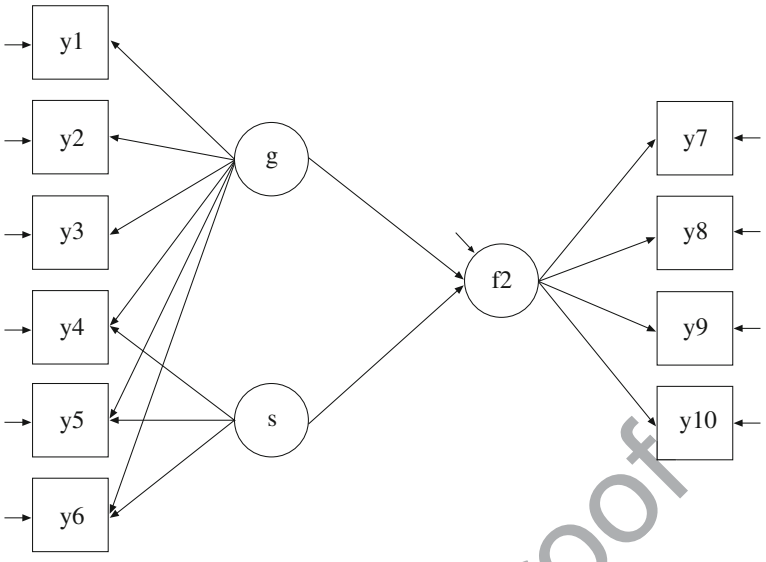

Fig. 12.2 Structural equation model with selection

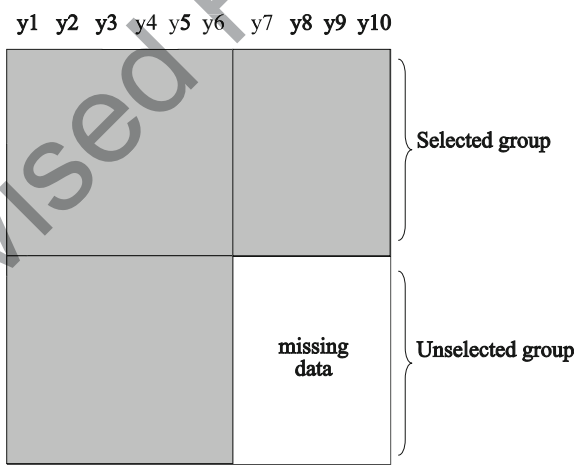

Fig. 12.3 Data structure with selection (non-shaded area represents missing data)

12.4 Using Listwise Deletion

Using listwise deletion, the sample of selected subjects is analyzed with respect to the model for y_1–y_{10} using the ML fitting function. Note that this does not give ML estimates of the parameters in the full sample. With 30 degrees of freedom the mean and variance of the likelihood-ratio χ^2 test are expected to be 30 and 60, but are

somewhat overestimated as 32.306 and 72.915 and the 5 % reject proportion obtains a somewhat too high value of 0.108. Still, this implies that the model would often not be rejected.

The results for the parameter estimates are shown in Tables 12.1 and 12.2. The first column shows the population values that were chosen and with which the data were generated. In terms of the selection instrument y_1–y_6, the factor loadings for the general factor g and the specific factor s are clearly misestimated as is seen in the Average column. The 95 % coverage column also shows large deviations from 95 % coverage. Standardized versions of the factor loadings for y_2 and y_5 are shown at the bottom of Table 12.2 as stdlam2 and stdlam5g, stdlam5s. This indicates that the variance explained by the general factor is underestimated and the variance explained by the specific factor is overestimated. This reflects the fact that the selection variable is most closely aligned with g given that selection is based on a sum of all the variables y_1–y_6.

The key parameters of the structural equation relating f2 to g and s show that the influence of g is underestimated and the influence of s is overestimated. It is seen that the variance for g is more strongly underestimated than the variance for s as is expected due to the selection being more closely aligned with g. Table 12.2 shows that this results in a standardized effect of g on f2 that is strongly underestimated as 0.4396 compared to the true value of 0.7. At the same time, the standardized effect of s on f2 is overestimated as 0.6664 instead of 0.506. That is, the relative importance of the two factors is reversed, distorting the true predictive value of the factors in the full population.

For the criterion instrument Table 12.1 shows that the unstandardized factor loadings are well estimated with good coverage. The standardized factor loading for y_8, listed as stdlam8, shows a slight overestimation, which is due to the variance of f2 being underestimated (see the vf2 entry).

The results of the Monte Carlo simulations can be explained via Pearson-Lawley formulas applied to factor analysis. The key results are discussed in Sect. 12.6, whereas Section? presents this in technical terms using matrix formulas.

12.5 Using ML Under MAR

Consider again the model of Fig. 12.2 and the data structure of Fig. 12.3 showing that there is missing data on y_7–y_{10} for subjects who are not selected. In the Monte Carlo study of the previous section, model estimation considered subjects in the selected group who have complete data on y_1–y_{10}. Using the same Monte-Carlo generated data, maximum-likelihood estimation is now applied under the MAR assumption. MAR is fulfilled because the missingness for y_7–y_{10} is determined by the variables y_1–y_6 which are observed with no missingness. Maximum-likelihood estimation uses all available data, that is, not only subjects in the selected group who have complete data on y_1–y_{10}, but also subjects in the unselected group who have data on only y_1–y_6.

Table 12.1 Results obtained by listwise deletion

	Estimates			S.E.	M.S.E.	95 %	% sig
	Population	Average	Std. dev.	Average		Cover	Coeff
g BY							
y1	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
y2	0.800	0.8451	0.0959	0.0944	0.0112	0.958	1.000
y3	0.700	0.5687	0.0752	0.0740	0.0229	0.546	1.000
y4	0.800	0.3624	0.0773	0.0770	0.1975	0.000	1.000
y5	0.700	0.4458	0.0620	0.0638	0.0685	0.040	1.000
y6	0.600	0.2618	0.0638	0.0657	0.1185	0.004	0.978
s BY							
y4	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
y5	0.800	0.8385	0.0889	0.0880	0.0094	0.968	1.000
y6	0.700	0.6785	0.0934	0.0868	0.0092	0.920	1.000
f2 BY							
y7	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
y8	0.800	0.8017	0.0447	0.0454	0.0020	0.940	1.000
y9	0.700	0.7029	0.0428	0.0418	0.0018	0.934	1.000
y10	0.600	0.6007	0.0394	0.0385	0.0016	0.952	1.000
f2 ON							
g	0.700	0.5754	0.0745	0.0748	0.0211	0.592	1.000
s	0.800	0.9566	0.1071	0.1053	0.0360	0.738	1.000
g WITH							
s	0.000	0.0000	0.0000	0.0000	0.0000	1.000	0.000
Intercepts							
y1	0.000	0.7932	0.0314	0.0311	0.6301	0.000	1.000
y2	0.000	0.6202	0.0266	0.0255	0.3854	0.000	1.000
y3	0.000	0.6137	0.0292	0.0302	0.3774	0.000	1.000
y4	0.000	0.8351	0.0334	0.0338	0.6986	0.000	1.000
y5	0.000	0.6745	0.0257	0.0263	0.4555	0.000	1.000
y6	0.000	0.6356	0.0274	0.0292	0.4047	0.000	1.000
y7	0.000	0.6170	0.0367	0.0363	0.3820	0.000	1.000
y8	0.000	0.4921	0.0278	0.0282	0.2430	0.000	1.000
y9	0.000	0.4305	0.0259	0.0265	0.1860	0.000	1.000
y10	0.000	0.3691	0.0275	0.0249	0.1370	0.000	1.000
Variances							
g	1.000	0.3689	0.0546	0.0520	0.4013	0.000	1.000
s	0.400	0.3075	0.0487	0.0481	0.0109	0.492	1.000

(continued)

Table 12.1 (continued)

	Estimates			S.E.	M.S.E.	95 %	% sig
	Population	Average	Std. dev.	Average		Cover	Coeff
Residual variances							
y1	0.600	0.6003	0.0494	0.0479	0.0024	0.946	1.000
y2	0.400	0.3913	0.0340	0.0329	0.0012	0.942	1.000
y3	0.800	0.7925	0.0394	0.0391	0.0016	0.930	1.000
y4	0.800	0.7838	0.0455	0.0463	0.0023	0.942	1.000
y5	0.400	0.4064	0.0272	0.0270	0.0008	0.950	1.000
y6	0.700	0.6883	0.0364	0.0351	0.0015	0.932	1.000
y7	0.700	0.7003	0.0385	0.0410	0.0015	0.968	1.000
y8	0.400	0.3982	0.0240	0.0243	0.0006	0.948	1.000
y9	0.400	0.3974	0.0233	0.0223	0.0005	0.934	1.000
y10	0.400	0.3991	0.0212	0.0209	0.0005	0.952	1.000
f2	0.254	0.2237	0.0350	0.0358	0.0021	0.858	1.000

Table 12.2 Standardized results obtained by listwise deletion

	Estimates			S.E.	M.S.E.	95 %	% sig
	Population	Average	Std. dev.	Average		Cover	Coeff
New/additional parameters							
vy2	1.200	0.6512	0.0292	0.0291	0.3021	0.000	1.000
vy5	1.146	0.6930	0.0312	0.0310	0.2062	0.000	1.000
vf2	1.000	0.6230	0.0556	0.0567	0.1452	0.000	1.000
vy8	1.040	0.6228	0.0495	0.0517	0.1765	0.000	1.000
stdf2ong	0.700	0.4396	0.0422	0.0423	0.0696	0.000	1.000
stdf2ons	0.506	0.6664	0.0413	0.0413	0.0274	0.036	1.000
stdlam2	0.800	0.6302	0.0393	0.0388	0.0304	0.006	1.000
stdlam5g	0.654	0.3227	0.0375	0.0384	0.1112	0.000	1.000
stdlam5s	0.473	0.5534	0.0357	0.0362	0.0077	0.388	1.000
stdlam8	0.784	0.8007	0.0256	0.0256	0.0009	0.896	1.000

The Monte Carlo results for ML are as follows. The likelihood-ratio χ^2 test performs well. With 30 degrees of freedom χ^2 has mean 29.933, variance 61.528, and 5 % reject proportion 0.050, which are all close to the expected values. The parameter estimation works very well as shown in Table 12.3. The ML approach of also using the information on y_1–y_6 for those not selected produces estimates close to the true values for not only the y_1–y_6 part of the model but for the whole model.

As a minor detail, it may be noted that the f2 factor loadings of Table 12.3 have smaller standard errors than those in Table 12.1 using the selected group only. This reflects the smaller sample size when using only the selected group.

Table 12.3 Maximum-likelihood results assuming MAR

	Estimates			S.E.	M.S.E.	95 %	% sig
	Population	Average	Std. dev.	Average		Cover	Coeff
g BY							
y1	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
y2	0.800	0.7983	0.0213	0.0214	0.0005	0.948	1.000
y3	0.700	0.6992	0.0193	0.0199	0.0004	0.952	1.000
y4	0.800	0.7997	0.0262	0.0262	0.0007	0.952	1.000
y5	0.700	0.7011	0.0230	0.0235	0.0005	0.960	1.000
y6	0.600	0.6012	0.0216	0.0217	0.0005	0.948	0.978
s BY							
y4	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
y5	0.800	0.7981	0.0475	0.0466	0.0023	0.938	1.000
y6	0.700	0.6993	0.0441	0.0425	0.0019	0.950	1.000
f2 BY							
y7	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
y8	0.800	0.8004	0.0386	0.0401	0.0015	0.944	1.000
y9	0.700	0.7020	0.0390	0.0375	0.0015	0.926	1.000
y10	0.600	0.5996	0.0356	0.0352	0.0013	0.942	1.000
f2 ON							
g	0.700	0.6981	0.0413	0.0449	0.0017	0.976	1.000
s	0.800	0.8011	0.0612	0.0614	0.0037	0.956	1.000
g WITH							
s	0.000	0.0000	0.0000	0.0000	0.0000	1.000	0.000
Intercepts							
y1	0.000	0.0001	0.0259	0.0265	0.0007	0.964	0.036
y2	0.000	0.0002	0.0234	0.0228	0.0005	0.952	0.048
y3	0.000	0.0013	0.0203	0.0211	0.0004	0.954	0.046
y4	0.000	0.0000	0.0276	0.0268	0.0008	0.942	0.058
y5	0.000	0.0008	0.0242	0.0240	0.0006	0.946	0.054
y6	0.000	0.0010	0.0210	0.0219	0.0004	0.960	0.040
y7	0.000	0.0054	0.0459	0.0467	0.0021	0.958	0.042
y8	0.000	0.0036	0.0392	0.0407	0.0015	0.956	0.044
y9	0.000	0.0019	0.0379	0.0379	0.0014	0.958	0.042
y10	0.000	0.0029	0.0358	0.0353	0.0013	0.940	0.060
Variances							
g	1.000	1.0032	0.0454	0.0457	0.0021	0.954	1.000
s	0.400	0.4013	0.0319	0.0328	0.0010	0.968	1.000

(continued)

Table 12.3 (continued)

	Estimates			S.E.	M.S.E.	95 %	% sig
	Population	Average	Std. dev.	Average		Cover	Coeff
Residual variances							
y1	0.400	0.3990	0.0213	0.0212	0.0005	0.956	1.000
y2	0.400	0.3992	0.0181	0.0170	0.0003	0.922	1.000
y3	0.400	0.4001	0.0155	0.0157	0.0002	0.948	1.000
y4	0.400	0.3984	0.0225	0.0229	0.0005	0.962	1.000
y5	0.400	0.4007	0.0178	0.0179	0.0003	0.954	1.000
y6	0.400	0.3994	0.0167	0.0162	0.0003	0.950	1.000
y7	0.400	0.4006	0.0262	0.0274	0.0007	0.970	1.000
y8	0.400	0.3981	0.0230	0.0229	0.0005	0.954	1.000
y9	0.400	0.3974	0.0224	0.0214	0.0005	0.940	1.000
y10	0.400	0.3993	0.0208	0.0203	0.0004	0.954	1.000
f2	0.254	0.2518	0.0278	0.0283	0.0008	0.944	1.000
New/additional parameters							
vy2	1.040	1.0375	0.0327	0.0328	0.0011	0.964	1.000
vy5	1.146	1.1480	0.0360	0.0363	0.0013	0.952	1.000
vf2	1.000	0.9995	0.0934	0.0977	0.0087	0.958	1.000
vy8	1.040	1.0386	0.0561	0.0587	0.0031	0.962	1.000
stdf2ONg	0.700	0.6995	0.0258	0.0267	0.0007	0.964	1.000
stdf2ONs	0.506	0.5072	0.0336	0.0343	0.0011	0.952	1.000
stdlam2	0.784	0.7839	0.0137	0.0124	0.0002	0.904	1.000
stdlam5 g	0.654	0.6545	0.0156	0.0158	0.0002	0.956	1.000
stdlam5 s	0.473	0.4709	0.0231	0.0245	0.0005	0.970	1.000
stdlam8	0.784	0.7843	0.0189	0.0190	0.0004	0.932	1.000

12.6 Pearson-Lawley Selection Formulas

In the regression example there is one selection variable and it is identical to x. In general, the selection variable need not be the same as x, need not be an observed variable, and need not be a single variable. The Pearson-Lawley formulas assume linear, homoscedastic regression of a set of continuous analysis variables on a set of selection variables. Normal distributions are not assumed. The general Pearson-Lawley selection formulas can be used for deriving means, variances, and covariances for the full population given values of the selected population and vice versa.

Going from the full to a selected population, the means, variances, and covariances of the analysis variables in the selected population are obtained from (1) the means, variances, and covariances of the selection variables in the selected and full population; (2) the covariances of the analysis and selection variables in the full population; and (3) the means, variances, and covariances of the analysis variables in the full population.

Going from a selected to the full population, the means, variances, and covariances of the analysis variables in the full population are obtained from (1) the means, variances, and covariances of the selection variables in the selected and full population; (2) the covariances of the selection and analysis variables in the selected population; and (3) the means, variances, and covariances of the analysis variables in the selected population.

12.7 Pearson-Lawley and Factorial Invariance

As pointed out in Meredith (1964), see also Olsson (1978) and Muthén and Jöreskog (1983), a factor model for a certain population also holds in a selected subpopulation if selection takes place on variables related to the factors and not directly related to the factor indicators. This is in line with regression where selection on x does not change the regression parameters, but selection on y does (Muthén and Jöreskog 1983).

When selection is related to only the factors, the full population factor loadings, factor indicator intercepts, and factor indicator residual variances are not affected by selection but are the same in the selected population. This is a rationale for assuming scalar measurement invariance in multiple-group modeling. The selection effect is absorbed into the factor means and the factor covariance matrix (see, e.g., Muthén and Jöreskog 1983, p. 367; Muthén et al. 1987, p. 440). Consider, for example, the case of a gender covariate influencing the factors. In a two-group analysis based on gender one should therefore expect full measurement invariance. In contrast, consider the gender covariate influencing factor indicators directly, where the direct effects imply that the means of the factor indicators vary across gender not only as a function of the factor mean varying across gender. In this case, selection on gender implies selection on factor indicators and one should not expect full measurement invariance. When selection is directly related to the factor indicators, the factor model does not hold in the selected subpopulation but is distorted as shown in Muthén (1989, p. 83) and illustrated in the Monte Carlo simulation.

For Fig. 12.2 model used in the Monte Carlo study, the factor model for y_1–y_6 is distorted because of selection on the factor indicators. The factor model for y_7–y_{10} is not distorted, however, because the selection is indirect via the factor f2 given that y_1–y_6 do not influence y_7–y_{10} directly. The next section discusses an approach that gives correct maximum-likelihood estimates under this type of selection.

12.8 Predictive Validity of Factors

Structural equation models are useful in predictive validity studies given that factors playing different roles in the test performance can be isolated and used as predictors of criterion outcomes. The use of a bi-factor model such as Fig. 12.2 is studied e.g.

in Gustafsson and Balke (1993), arguing for the value of using both a general and specific factors. While several previous studies indicate that not much increase in predictive power is to be gained from using a differentiated set of ability dimensions, as compared to an undifferentiated composite score (see, e.g. Schmidt and Hunter 1981), Gustafsson and Balke (1993) demonstrate that a bi-factor, orthogonal factor model may bring out a more differentiated pattern of relations between predictors and criteria, and particularly so if a latent variable model is used also for the criterion variables.

Muthén and Hsu (1993) study selection and predictive validity for structural equation models such as those used in Gustafsson and Balke (1993) One of their approaches uses factor scores based on the parameters from the factor model for the predictors estimated from a random sample of the full population. This corresponds to using all subjects of Fig. 12.3. In the case of a random sample, that is, no selection, it is known (Tucker 1971) that with factor score estimated by the regression method, consistent estimates are obtained for the regression of a dependent variable on the estimated factor scores. Although the factor scores have biases, the factor covariance bias and the bias in the covariances of the factors and a dependent variable cancel out in the regression of a dependent variable on the estimated factor scores. In the current case of selection, Muthén and Hsu (1993, pp. 261–262) use Pearson-Lawley selection formulas to show that when a sum of the factor indicators is used as a selection variable, the regression of a dependent variable on the estimated factor scores in the selected group also gives unbiased structural coefficients.

12.9 Selection Based on Factors: Predictive Validity of Admission Tests in the U.S. Military

Given a model such as Fig. 12.2 it is of interest to select subjects based on the factor values instead of a sum of the factor indicators. Muthén and Gustafsson (1994) compare selection based on factors with the conventional selection based on sums of factor indicators used for admission into the U.S. military. Hands-on job performance for nine U.S. army jobs is related to the standard set of ten ASVAB tests as well as twelve experimental tests added to the ASVAB. A bi-factor model is considered for the total number of 22 tests.

A first complication is that it is not known who among the applicant sample was selected and who was not. This means that the data are not structured as in Fig. 12.3 because information on y_1–y_6 for an unselected group is not available. A second complication is that data for the 22 tests are not available for the unselected, applicant group. Only the ten ASVAB tests are available for the applicant group and the twelve experimental tests are only available for the selected, matriculant group. These two complications are resolved by using Pearson-Lawley adjustments in combination with the factor score approach as follows.

As a first step, Pearson-Lawley adjustment is made to the 22 × 22 sample covariance matrix for the nine jobs in the selected, matriculant group to obtain an estimate of the covariance matrix for the unselected, applicant group. In this adjustment the ten ASVAB tests are used as selection variables given that ASVAB is the standard selection instrument into the military. A 10 × 10 ASVAB covariance matrix is used for a reference group of 650,278 applicants. A bi-factor factor model is then applied to the 22 × 22 adjusted covariance matrix and estimated factor scores computed for the selected group in the nine army jobs. The criterion variable of hands-on job performance is then regressed on the estimated factor scores to give unbiased regression estimates in line with Muthén and Hsu (1993). Hsu (1995) shows that standard errors for these regression estimates are well approximated at moderate sample sizes even though the factor score estimation assumes no sampling error in the factor model parameters. Muthén and Gustafsson (1994) show that different profiles of selected subjects are obtained using the factor-based selection versus using the conventional selection. They also argue that the assessment of incremental predictive validity of new tests is better done using a factor model.

12.10 Swedish Military Enlistment Example

A special application of the maximum-likelihood approach under MAR is used for Swedish military enlistment data in Muthén et al. (1994). Enlistment data collection includes: a cognitive test battery; a psychologist's rating of ability to handle strenuous situations; education; medical, and physical tests; and a psychologist's rating of the suitability for being an officer. Performance is measured as two supervisor ratings at the end of the training. Here, the missing data structure features three missing data patterns. Only the individuals scoring in the top 60–70 % of the cognitive test are evaluated for their suitability for being an officer, and performance is only measured for individuals selected as officers. Selection as officer is determined by several other factors than those determining the missing data patterns.

Muthén et al. (1994) use data on the performance of a select group of officers in charge of large units. Because individuals are not followed longitudinally the data come from two sources, a criterion sample of 1208 graduating officers and an enlistment sample. The enlistment sample is created as a random subsample of individuals known to have been selected as officers from the three years that the criterion sample officers were most likely tested. A sample size corresponding to the known selection ratio is chosen so that the maximum-likelihood procedure has the proper ratio of selected and non-selected individuals. A latent variable model is formulated with three latent variable constructs for the four cognitive tests, one construct for the psychologist's ratings, and one construct for the supervisor ratings. In a preliminary analysis, logistic regression is carried out to study predictors of being selected as an officer. In addition to the variables listed above, the location of the enlistment office and the time between the enlistment testing and service are

found important and are included in the final latent variable model to avoid selection biases. A useful finding for modifying the selection procedure concerns the time between the enlistment testing and service. While increasing time has a negative effect on selection, it has a positive effect on performance as an officer, presumably due to an age advantage.

12.11 Conclusions

This chapter shows how analysis of data on selective subgroups can be used to draw inference to the full, unselected group. This uses Pearson-Lawley selection formulas which apply to not only regression analysis but also structural equation modeling. The chapter shows the connection with maximum-likelihood estimation with missing data assuming MAR versus using listwise deletion. Applications are discussed of selection into the military using factor analysis models for the variables used in the selection.

References

Dunbar, S. B., & Linn, R. L. (1991). Range restriction adjustments in the prediction of military job performance. In A. K. Wigdor & B. F. Green (Eds.), *Performance assessment for the workplace* (Vol. II. Technical issues, pp. 127–157). Washington, DC: National Academy Press.

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*, 407–434.

Hsu, J. W. Y. (1995). Sampling behaviour in estimating predictive validity in the context of selection and latent variable modelling: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 48*, 75–97.

Johnson, N. L., & Kotz, S. (1972). *Distributions in statistics: Continuous multivariate distributions*. Chichester: Wiley.

Lawley, D. (1943–1944). A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh: Proceedings, Section A*, *62*, 28–30.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*, 177–185.

Muthén, B. (1989). Factor structure in groups selected on observed scores. *British Journal of Mathematical and Statistical Psychology*, *42*, 81–90.

Muthén, B., & Gustafsson, J. E. (1994). *ASVAB-based job performance prediction and selection: Latent variable modeling versus regression analysis*. Technical report.

Muthén, B. & Hsu, J. W. Y. (1993). Selection and predictive validity with latent variable structures. *British Journal of Mathematical and Statistical Psychology*, *46*, 255–271.

Muthén, B., & Hsu, J. W. Y., Carlstedt, B., & Mardberg, B. (1994). *Predictive validity assessment of the Swedish military enlistment testing procedure using missing data and latent variable methods*. Technical report.

Muthén, B., & Jöreskog, K. (1983). Selectivity problems in quasi-experimental studies. *Evaluation Review*, *7*, 139–174.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431–462.
Olsson, U. (1978). *Selection bias in confirmatory factor analysis*. Uppsala, Sweden: University of Uppsala (Department of Statistics Research, Report No. 78-4).
Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI: On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, London, Series A, 200*, 1–66.
Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.
Tucker, L. R. (1971). Relations of factor score estimates to their use. *Psychometrika, 36*, 427–436.