14

Latent Growth Curve Models

An interesting structural equation model for fixed occasion panel data is the Latent Curve Model (LCM). This model has been applied mainly to developmental or growth data, hence the usual name Latent Growth Curve Model. In the latent curve model, the time variable is defined in the measurement model of the latent factors. For instance, in a linear growth model, consecutive measurements are modeled by a latent variable for the intercept of the growth curve, and a second latent variable for the slope of the curve.

Figure 2 shows the path diagram of a simple latent curve model for panel data with five occasions, and one time-independent explanatory variable Z. In Figure 13.1, Y_0, Y_1, Y_2, Y_4 and Y_4 are the observations of the response variable at the five time points. In the latent curve model, the expected score at time point zero is modeled by a latent *intercept* factor. The intercept is constant over time, which is modeled by constraining the loadings of all time points on the intercept factor equal to one. The latent slope factor is the slope of a linear curve, modeled by constraining the loadings of the five time points on this factor to be equal to 0, 1, 2, 3 and 4 respectively. Evidently, a quadratic trend would be specified by a third latent variable, with successive loadings constrained to be equal to 0, 1, 4, 9 and 16. What is not immediately obvious from the path diagram in Figure 13.1, is that the latent curve model includes the means of the variables and the factors in the model. As a consequence, the regression equations that predict the observed variables from the latent factors, depicted by the single-headed arrows towards the observed variables in Figure 13.1, also contain terms for the intercept.

In the latent curve model, the intercepts of the response variable at the five time points are constrained to zero, and as a result, the mean of the intercept factor is an estimate of the common intercept.

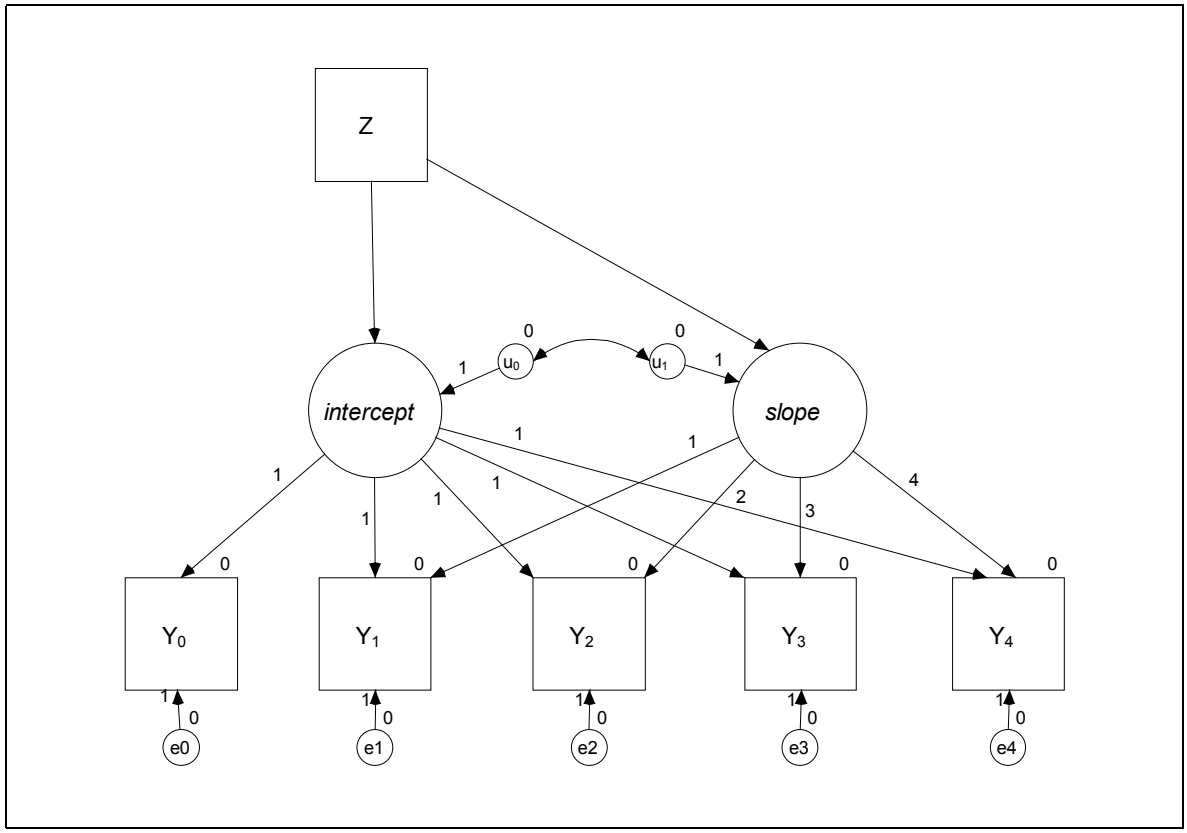

Figure 14.1 Latent Curve Model for five occasions

The successive loadings for the slope factor define the slope as the linear trend over time (note that, as is common in SEM, the first path from the slope factor to Y_0, which is equal to zero, is omitted from the diagram). The mean of the slope factor is an estimate of the common slope (c.f. Meredith & Tisak, 1990; Muthén, 1991; Willet & Sayer, 1994, Duncan & Duncan, 1995, Maccallum & Kim, 2000). Individual deviations from the common intercept are modeled by the variance of the intercept factor, and individual deviations in the slope of the curve are modeled by the variance of the slope factor. Both the intercept and slope factor can be modeled by a path model including explanatory variables, in our example the one explanatory variable Z

The latent curve model is a random coefficient model for change over time, equivalent to the multilevel regression model for longitudinal data that is described in Chapter 5. To clarify the relationship between the two models, we write the equations for both specifications. In the multilevel linear growth model, the model described by Figure 13.1 can be expressed as a multilevel regression model, with at the lowest level, the occasion level:

$$Y_{tj} = \pi_{0i} + \pi_{1i}T_{ti} + e_{ti}., \qquad (14.1)$$

where T_{ti} is an indicator for the occasions, which is set to 0, 1, 2, 3, 4 to indicate the five occasions. At the second level, the individual level, we have

$$\pi_{0i} = \beta_{00} + \beta_{01}Z_i + u_{0i}, \quad (14.2)$$
$$\pi_{1j} = \beta_{10} + \beta_{11}Z_i + u_{1i}, \quad (14.3)$$

By substitution, we get the single equation model:

$$Y_{tj} = \beta_{00} + \beta_{10}T_{ti} + \beta_{01}Z_i + \beta_{11}Z_iT_{ti} + u_{1i}T_{ti} + u_{0i} + e_{ti}. \quad (14.4)$$

In a typical SEM notation, we can express the path model in Figure 13.1 as:

$$Y_{ti} = \lambda_{0t}\,\text{intercept}_i + \lambda_{1t}\,\text{slope}_i + e_{ti}\,, \quad (14.5)$$

where λ_{0t} are the factor loadings for the intercept factor, and λ_{1t} are the factor loadings for the slope factor.

Note the similarity between the equations (14.5) and (14.1). In both cases, we model an outcome variable than varies across times *t* and individuals *i*. In equation (14.1), we have the intercept term π_{0i}, which varies across individuals. In equation (14.5), we have a latent intercept factor, which varies across individuals, and is multiplied by the factor loadings λ_{0t} to predict the Y_{tj}. Since the factor loadings λ_{0t} are all equal to one, it can be left out of equation (14.5), and we see that the intercept factor in equation (14.5) is indeed equivalent to the regression coefficient π_{0i} in equation (14.1). Next, in equation (14.1), we have the slope term π_{1i}, which varies across individuals, and is multiplied by the 0, …, 4 values for the occasion indicator T_{ti}. In equation (14.5), we have a latent slope factor, which varies across individuals, and gets multiplied by the factor loadings λ_{0t} to predict the Y_{tj}. Since the factor loadings λ_{1t} are set to 0, …, 4, we see that the slope factor in equation (14.5) is indeed equivalent to the regression coefficient π_{1i} in equation (14.1). Therefore, the fixed factor loadings for the slope factor play the role of the time variable T_{ti} in the multilevel regression model, and the slope factor plays the role of the slope coefficient π_{1i} in the multilevel regression model.

In a manner completely analogous to the second level equations (14.2) and (14.3) in the multilevel regression model, we can predict the intercept and the slope factor using the time-independent variable *Z*. For these equations, using for consistency the same symbols for the regression coefficients, we have

$$intercept_i = \beta_{00} + \beta_{01}Z_i + u_{0i}, \quad (14.6)$$

$$slope_j = \beta_{10} + \beta_{11} Z_i + u_{1i}, \tag{14.7}$$

which lead to a combined equation

$$Y_{ti} = \beta_{00} + \beta_{10}\,\lambda_{1t} + \beta_{01} Z_i + \beta_{11} Z_i\,\lambda_{1t} + u_{1i}\,\lambda_{1t} + u_{0i} + e_{ti} \tag{14.8}$$

Keeping in mind that the factor loadings 0, …, 4 in λ_{1t} play the role of the occasion indicator variable in T_t, we see that the multilevel regression model and the latent curve model are indeed highly similar. The only difference so far is that multilevel regression analysis generally assumes one common variance for the lowest level errors e_{ti}, while structural equation analysis typically estimates different residual error variances for all observed variables. However, if we impose a constraint on the latent curve model, that the variances for e_0, …, e_4 are all equal, we have indeed the same model. Full maximum likelihood estimation, using either approach, should give essentially the same results.

14.1 EXAMPLE OF LATENT CURVE MODELING

The longitudinal *GPA* data from Chapter five are used again, with a standard latent curve model as in Figure 13.1 applied to the data. The example data are a longitudinal data set, with longitudinal data from 200 college students. The students' Grade Point Average (GPA) has been recorded for six successive semesters. At the same time, it was recorded whether the student held a job in that semester, and for how many hours. This is recorded in a variable 'job' (with categories 0=no job, 1=1-2 hours, 2=3-5 hours, 3=5-10 hours, 4=more than 10 hours), which for the purpose of this example is treated as an interval level variable. In this example, we also use the student variables high school GPA and sex (1=male, 2=female).

In a statistical package such as SPSS or SAS, these data are typically stored with the students defining the cases, and the repeated measurements as a series of variables, such as GPA1, GPA2, …, GPA6, and JOB1, JOB2, …, JOB6. As explained in Chapter Five, most multilevel regression software needs a different data structure. Latent curve analysis views the successive time point as different variables, and thus we can use such a data file as it is. We start with a model that includes only the linear trend over time. Figure 13.2 shows the path diagram for this model.

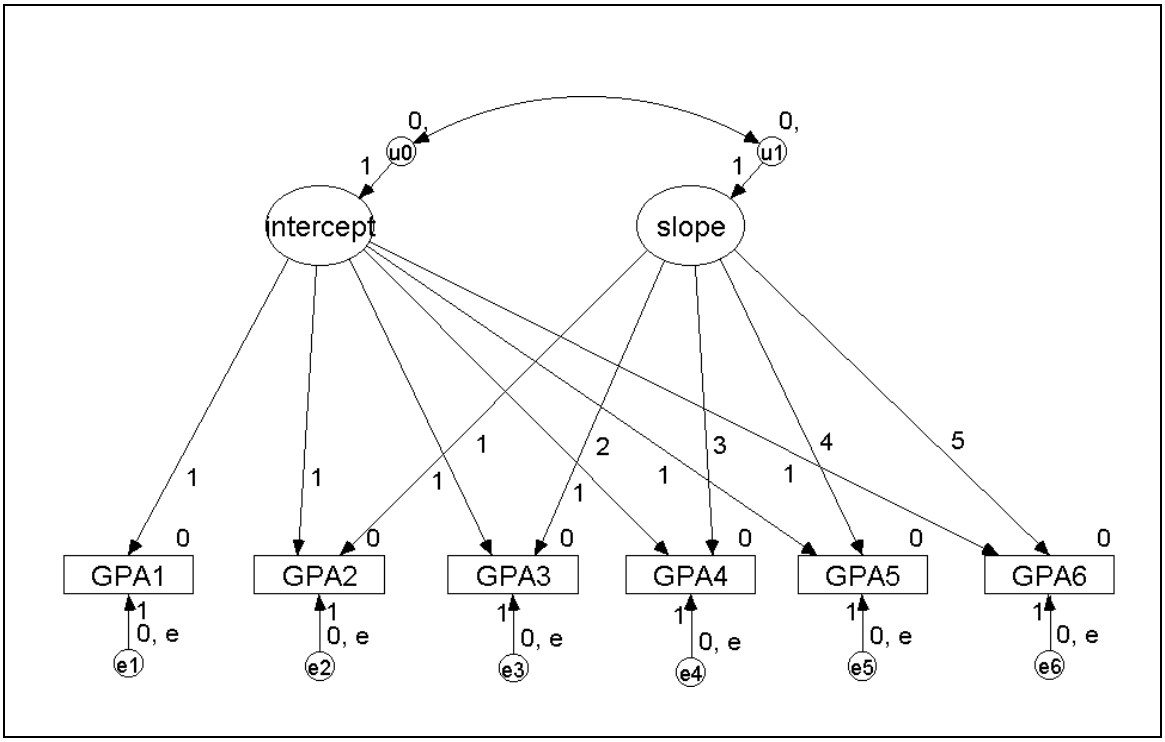

Figure 14.2 Path diagram for linear model GPA example data

The model in Figure 14.2 is equivalent to a multilevel regression model with a linear predictor coded 1, …, 6 for the successive occasions, and a random intercept and slope on the student level. To make the models completely equivalent, the error variances of the residual errors e_1, …, e_6 for the successive occasions are all constrained to be equal to e. The means of the intercept and slope factor are freely estimated, all other means and intercepts in the model are constrained to zero. The mean of the intercept is estimated as 2.60, and the mean of the slope is estimated as 0.11. This is identical to the estimates in the (fixed effects) multilevel regression model in Table 5.3 in Chapter Five.

For simplicity, we omit the time varying *JOB* variable for the moment, and start with specifying a latent curve model using only the six *GPA* scores, and the time-independent (student level) variables *high school GPA* and *student sex*. The path diagram, including the unstandardized parameter estimates, is shown in Figure 14.3.

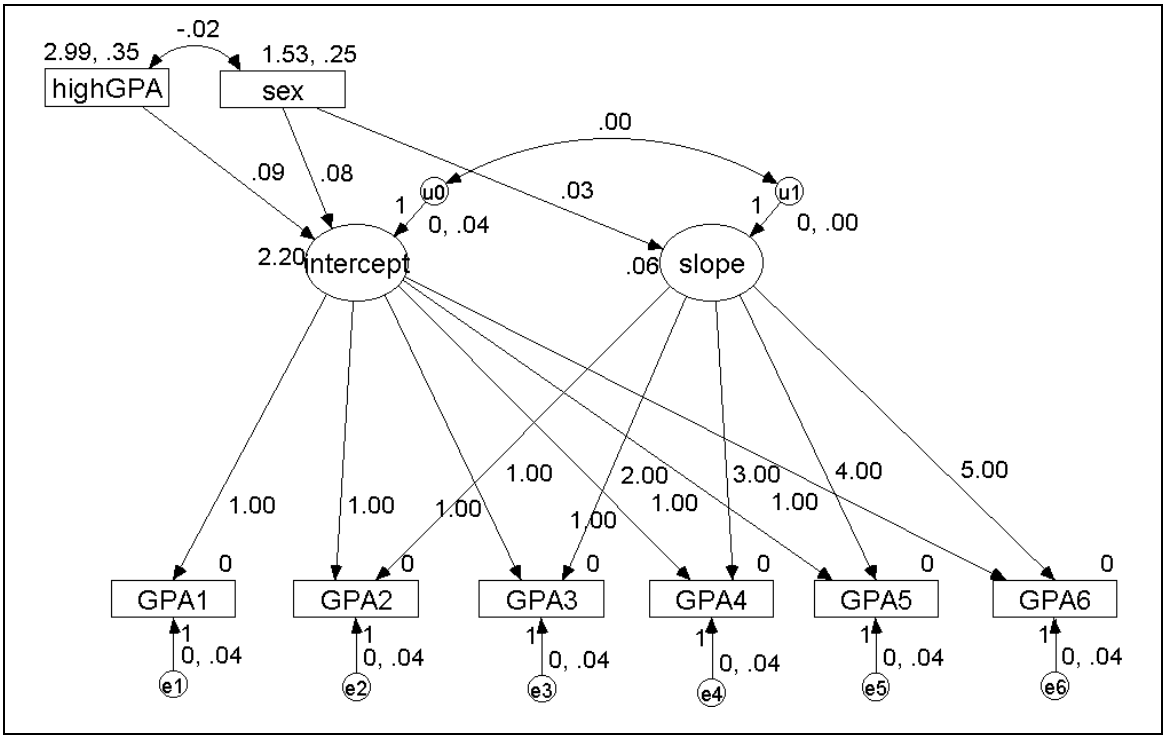

Figure 14.3 Path diagram and parameter estimates for linear curve model with two predictors

In the path diagram we see that in this model, which includes an interaction between the slope of the linear development over time and the student's sex, the average slope over time is 0.06. The slope variance in the figure is given in two decimals as 0.00, in the text output it is given as 0.004, with standard error 0.001. This is identical to the estimates in the similar multilevel regression model presented in Table 5.4 in Chapter Five.

The SEM analysis of the latent curve model gives us some information that is not available in the equivalent multilevel analyses. The models depicted in Figures 14.2 and 14.3 do not describe the data well. The model in Figure 14.2 has a chi-square of 190.8 (df=21, p<0.001) and a RMSEA fit index of 0.20, and the model in Figure 14.3 has a chi-square of 195.6 (df=30, p<0.001) and a RMSEA fit index of 0.17.[1] The SEM analysis also provides us with diagnostic information of the locus of the fit problem. The program output contains so called *modification indices* that signify constraints that decrease the fit of the model. All large modification indices indicate that the constraint of equal error variances for the residual errors e_1, …, e_6 does not fit well, and that the implicit constraint of no correlations between the residual errors e_1, …, e_6 does not fit well either. Presumably, the multilevel regression models presented in Chapter

[1] The GFI, CFI and TLI fit indices are problematic when the model includes means, because it is not clear what a proper null model would be for the means. The RSMEA does not have this limitation. Usually, an RMSEA ≤0.05 is judged as indicating a satisfactory fit.

Five also have these problems. Since in Chapter Five we did not carry out a residuals analysis or some other procedure to check for model misspecifications, we do not have any information about model fit. In SEM, we do have such information. If we remove the equality constraint on the residual errors, the model fit becomes much better, as indicated by a chi-square of 47.8 (df=25, p=0.01) and an RMSEA fit index of 0.07. Allowing correlated errors between the two first measurement occasions improves the fit to a chi-square of 42.7 (df=24, p=0.01) and an RMSEA of 0.06. Since the other estimates do not change much as a consequence of these modifications, the last model is accepted.

To bring the time varying variable *job status* into the model, we have several choices. Equivalent to the multilevel regression models for these data, which are treated in Chapter Five, we can add the variables job_1, …, job_6 as explanatory variables to the model. These predict the outcomes GPA_1, …, GPA_6, and since the multilevel regression model estimates only one single regression for the effect of *job status* on *GPA*, we must add equality constraints for these regression coefficients. The path diagram for this model is given in Figure 13.5.

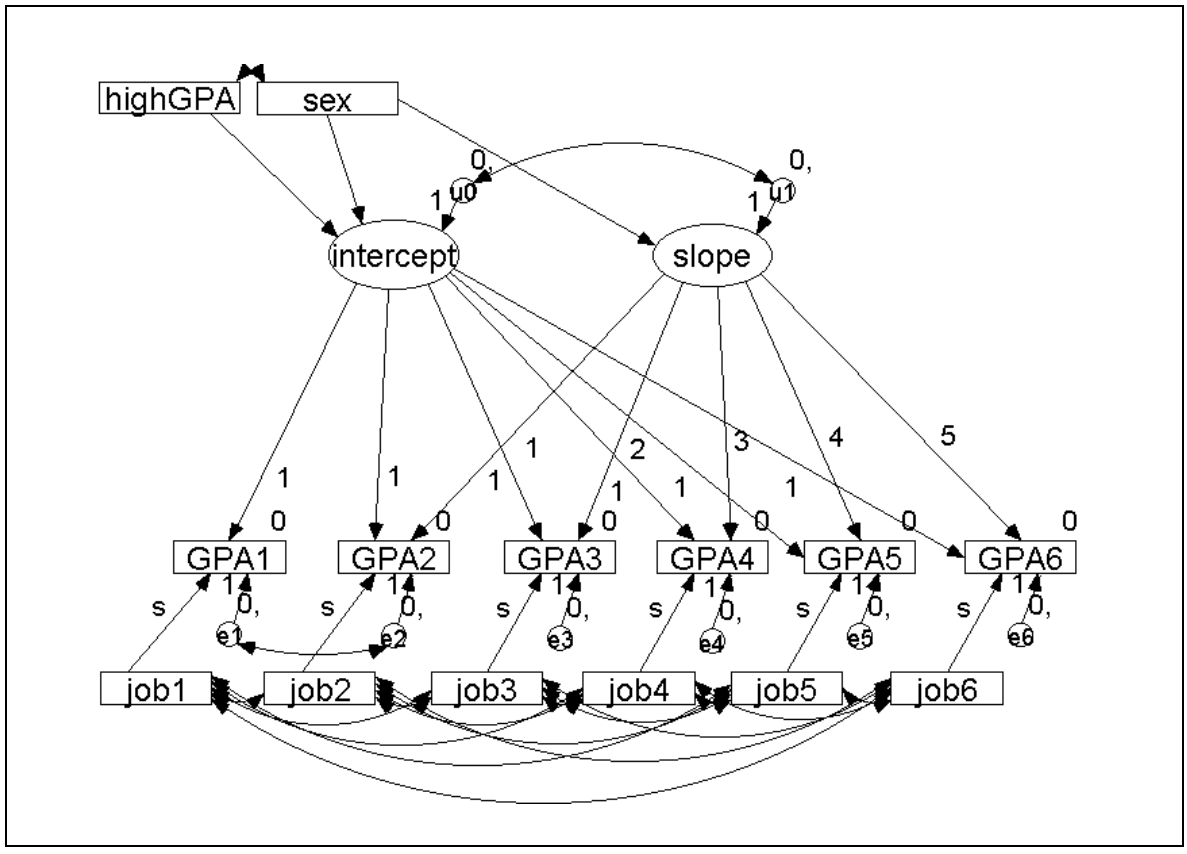

Figure 14.5 Path diagram for GPA example, including effects for job status

The common regression coefficient for *job status* on the *GPA* is estimates as –0.12 (s.e. 0.01), which is close to the multilevel regression estimates in Table 5.4. However, the model including all the job status variables does not fit well, with a chi-square of 202.1 (df=71, p<0.001) and an RMSEA of 0.10. There are no large modification indices, which indicates that there exists no single model modification, which substantially improves the model. We probably need many small modifications to make the model fit better.

An advantage of latent curve analysis is that it can be used to analyze more complex structures. For instance, we may attempt to model the changes in hours spend on a job using a second latent curve model. The path diagram for the latent curve model for *job status* at the six time points is presented in Figure 14.6.

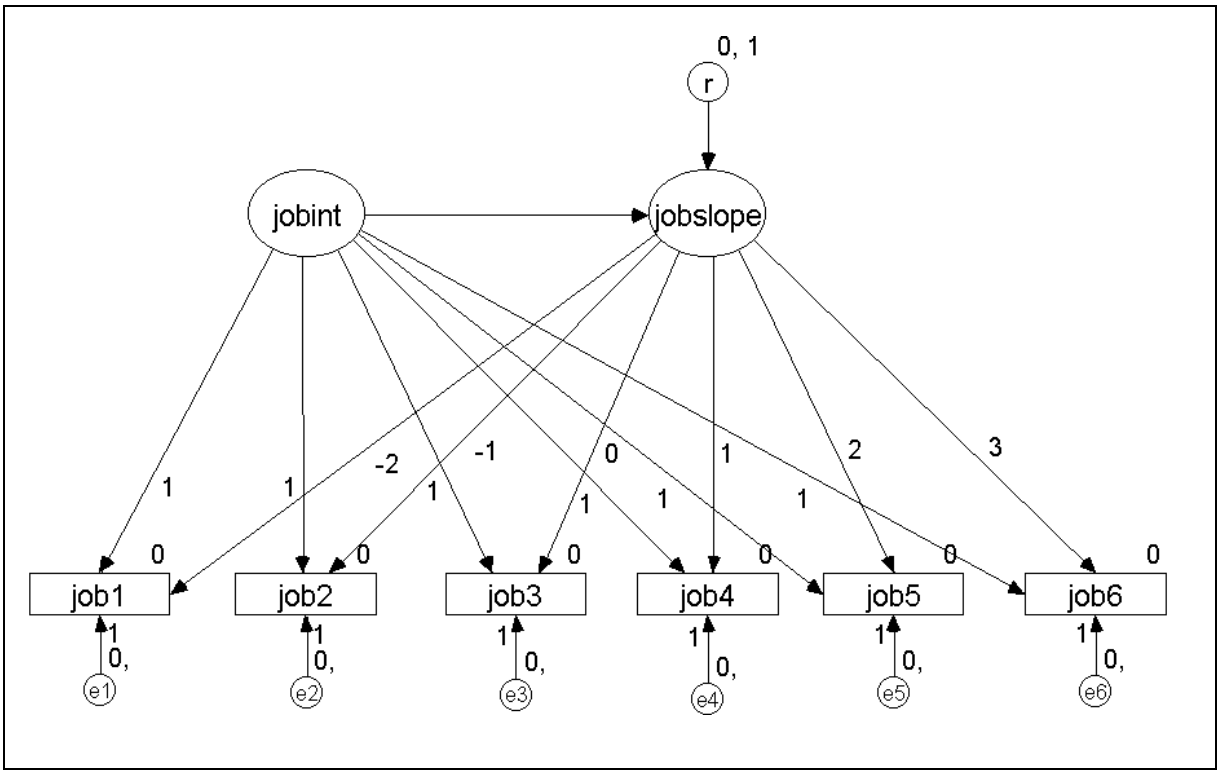

Figure 14.6 Latent curve model for *job status*

Figure 14.6 has some features that merit discussion. To ensure that the variances of the job intercept and slope factors are is positive, the variance is modeled by an error term *r* with the variance set at 1, and the path to *jobint* and *jobslope* estimated. The more usual procedure of setting the path at 1, and estimating the error variances, led to negative variance estimates.

The leads to a latent curve model that fits quite well, with a chi-square of 17.8 (*df*=17, *p*=0.40) and an RMSEA of 0.02. All estimates in this model are acceptable. The interesting feature of structural equation modeling is, that both models can be combined into one large model for change of both job status and GPA over time. Figure 14.7 shows one such model.

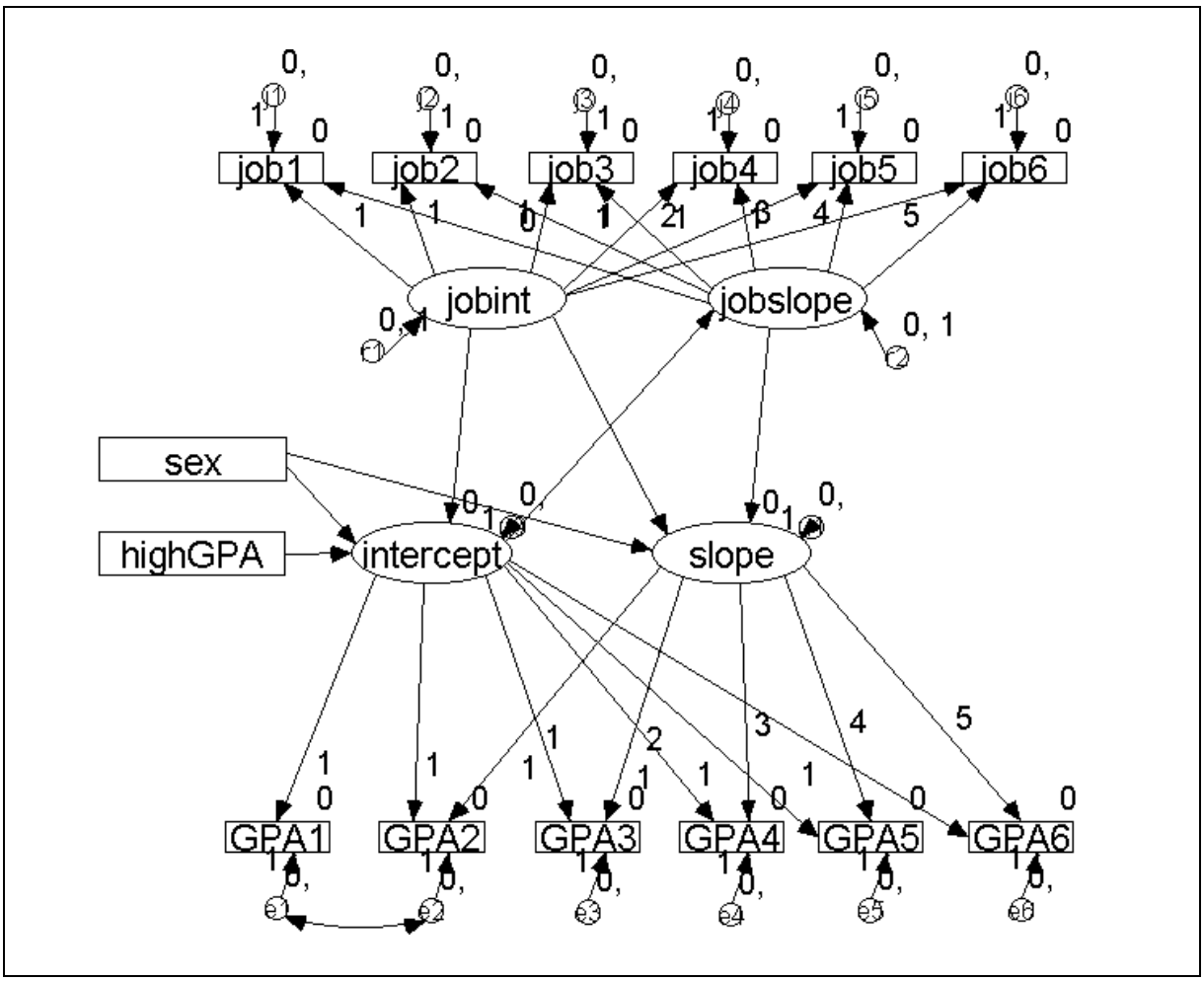

Figure 14.7 Path diagram for change in job status and GPA over time

The model depicted by the path diagram in Figure 13.7 shows a moderate fit. The chi-square is 166.0 (*df*=85, *p*<.001) and the RMSEA is 0.07. The AIC for the model in Figure 13.5, which is equivalent to a multilevel regression model, is 298.3. In comparison, the AIC for the model in Figure 13.5, which is *not* equivalent to a multilevel regression model, is 243.1. Although the complex latent curve model does not show an extremely good fit, it fits better than the multilevel regression model.

Figure 14.7 also shows that with complicated models with constraints on intercepts and variances, a path diagram becomes cluttered and difficult to read. Table 13.1 gives the estimates for the regression weights for the predictor variables *sex* and *high school GPA*, and the intercepts and slopes.

Table 14.1 Path coefficients for structural model in Figure 13.7

	Outcome		
Predictor	**job slope (s.e.)**	**GPA intercept (s.e.)**	**GPA slope (s.e.)**
sex		0.07 (.03)	0.02 (.01)
highGPA		0.07 (.02)	
job intercept		1.06 (.04)	0.03 (.01)
job slope			-0.46 (.11)
GPA intercept	-0.29 (.06)		

Figure 14.8 shows the same information, but now as standardized path coefficients with only the structural part of the path diagram shown.

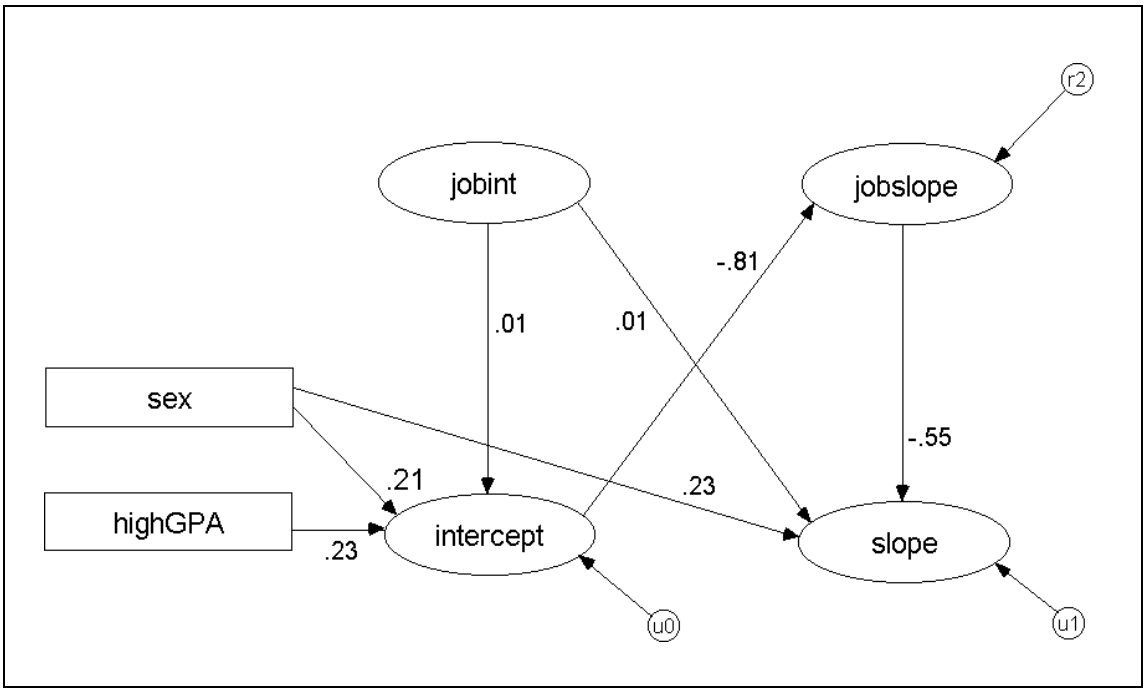

Figure 14.8 Estimated standardized path coefficients for structural model in Figure 14.7

Figure 14.8 shows results similar to the results obtained with the multilevel regression analyses in Chapter Five. Females have a higher GPA to begin with, and their GPA increases over the years at a faster rate than the male students do. The relations between the intercepts and slopes in Figure 14.8 show the mutual effects of changes over time in both job status and GPA. Initial job status has virtually no effect. Changes in job status, as reflected in the slope for job status, have a negative effect on the GPA. If the job status changes in the direction of spending more time on the job, the overall increase in GPA ends, and in fact can become negative. There is also an effect of initial GPA on job status: students with a high initial GPA increase their job workload less than other students.

14.2 SOME REMARKS ON LATENT CURVE MODELING

When multilevel regression analysis and latent curve modeling are applied to the same data set, the results are identical (cf. Chou, Bentler & Penz, 2000). Multilevel regression analysis has the advantage that adding one or several more levels is straightforward. When latent curve models are used, adding a group level is possible (cf. Muthén, 1997), but requires complicated program setups, or a specialized program (Muthén & Muthén, 1998).

As was remarked earlier in Chapter Five, multilevel regression copes automatically with missing data due to panel dropout. Since there is no requirement that each person has the same number of measurements, or even that the measures are taken at the same occasions, multilevel regression works very well on incomplete data. The latent curve model is a fixed occasions model. If different respondents are measured at different occasions, the latent curve model can deal with this only by specifying paths for all measurement occasions in the data set, and assuming that respondents have missing data at the occasions when they were not measured. Modern SEM software can estimate model parameters using incomplete data (Arbucle, 1996, Arbucle & Wothke, 1999, Muthén & Muthén, 1998), but when there are many and varying time points, the setup becomes complicated, and the estimation procedure may have convergence problems.

Latent curve models, on the other hand, have the advantage that it is straightforward to embed them in more complex path models. For instance, in latent growth methodology, it is simple to specify a path model where the slope factor is itself a predictor of some outcome. This represents a hypothesis that the rate of change is a predictor of some outcome. An example of such a path model was given in the previous section, where the rate of change in the latent slope for job status is a predictor for the rate of change indicated by the GPA slope

factor. This kind of hypothesis cannot be modeled in standard multilevel analysis. It is also simple to allow for different errors or correlated errors over time, which is possible in multilevel regression analysis, but more difficult to set up in the current software.

Since in latent curve models the model for the change over time is embedded in a structural equation model, it is also easy to extend by adding a measurement model for the variable that is measured over time. That could be indicated by a set of observed variables, and the variable that is modeled using the latent curve defined by the intercept and slope factors is then itself a latent variable.

A last advantage of the latent curve model is that standard SEM software provides information on goodness-of-fit, and suggests model changes that improve the fit.