Chapter 7

The consequences of unobserved heterogeneity in a sequential logit model

7.1 Introduction

Many processes can be described as a nested sequence of decisions or steps. Consider the three following examples. Mare (1979, 1980, 1981) describes the process of attaining education as the result of a sequence of transitions between educational levels, for example: 1) whether to finish secondary education or to leave school with only primary education, and 2) whether or not to finish tertiary education given that one finished secondary education. O'Rand and Henretta (1982) describe the decision when to retire using the following sequence of decisions: 1) whether to retire before age 62 or later, and 2) whether to retire before age 64 or later given that one has not retired before age 62. Cragg and Uhler (1970) describe the demand for automobiles as the result of the following sequence of decisions: 1) whether or not to buy an automobile, 2) whether to add an automobile or to replace an automobile given that one decided to buy an automobile, 3) whether or not to sell an automobile or not given that one decided not to buy an automobile. An attractive model for these processes is to estimate a separate logistic regression for each step or decision. These steps or decisions are often called transitions. This model is known under a variety of names: sequential response model (Maddala, 1983), sequential logit model (Tutz, 1991), continuation ratio logit (Agresti, 2002), model for nested dichotomies (Fox, 1997), and the Mare model (Shavit and Blossfeld, 1993). This model has however been subject to an influential critique by Cameron and Heckman (1998). Their main point starts with the observation that the sequential logit model, like any other model, is a simplification of reality and will not include all variables that influence the probability of passing a transition. The presence of these unobserved variables is often called unobserved heterogeneity, and it will lead to biased estimates, even if these unobserved variables are not confounding variables. There are two mechanisms through which these unobserved non-confounding variables will influence the results. The first mechanism, which I will call the averaging mechanism, is based on the fact that leaving a variable out of the model means that one models the probability of passing a transition averaged over the variable that was left out. The effect of the remaining variables on

this average probability of passing a transition is not the same as the effect of these variables on the probability that an individual passes that transition, because the relationship between the variable left out of the model and the probability is non-linear (Neuhaus and Jewell, 1993; Cameron and Heckman, 1998; Allison, 1999). The second mechanism, which I will call the selection mechanism, is based on the fact that even if a variable is not a confounding variable at the initial transition because it is uncorrelated with any of the observed variables, it will become a confounding variable at the higher transitions because the respondents who are at risk of passing these higher transitions form a selected sub-sample of the original sample (Mare, 1980; Cameron and Heckman, 1998).

The aim of this chapter is to propose a sensitivity analysis with which one can investigate the consequences of unobserved non-confounding variables in a sequential logit model. This will be done by specifying a set of plausible scenarios concerning this unobserved variability and estimating the individual-level effects within each of these scenarios, thus creating a range of plausible values for the individual-level effects.

Any method for studying such individual-level effects will have to deal with the fact that it tries to control for variables that have not been observed. This is a problem that also occurs with other models that try to estimate causal effects (Holland, 1986). A common strategy in these causal models is to use information that might be available outside the data. The clearest example of this is the experiment in which one knows that the respondents have been randomly assigned to the treatment and the control group, and it is this information that is being used to control for any unobserved variables. Various variations on this strategy have been proposed for non-experimental settings (Morgan and Winship, 2007), for example one might know that a variable influences the main explanatory variable but not the outcome variable, in which case one can use this variable as an instrumental variable, or one might know that all variables influencing the main explanatory variable are present in the data, in which case one can use propensity score matching. An example of such a strategy that has been applied to the sequential logit model is the model by Mare (1993, 1994), who used the fact that siblings are likely to have a shared family background. If one has data on siblings, one can thus use this information for controlling for unobserved variables on the family level. Another example of this strategy is the model used by Holm and Jæger (2008), who use instrumental variables in a sequential probit model[1] to identify individual-level effects. The strength of this strategy depends on the strength of the information outside the data that is being used to identify the model. However, such external information is often not available. In those cases, one can still use these mod-

[1] The sequential probit model is similar to the sequential logit model except that the probit link function is used rather than the logit link function.

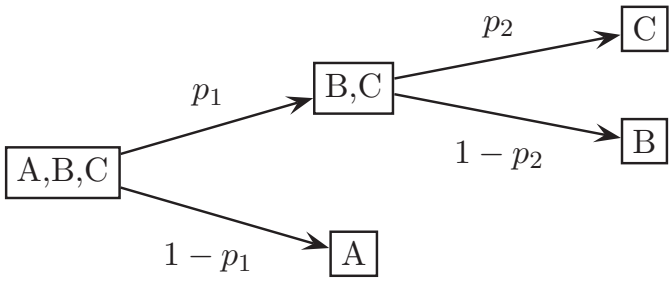
els, except that the identification is now solely based on untestable assumptions. This implies a subtle shift in the goal of the analysis: instead of trying to obtain an empirical estimate of a causal effect, one is now trying to predict what would happen if a certain scenario were true. This is not unreasonable: the causal effects are often the quantity of interest, and if it is not possible to estimate them, then the results of these scenarios are the next best thing. However, the modelling challenge now changes from making the best use of some information outside the data to finding the most informative comparison of scenarios. The goal of such an analysis is to find a plausible range of estimates of the causal effect and to assess how sensitive the conclusions are to changes in the assumptions (Rosenbaum and Rubin, 1983; Rosenbaum, 2002; DiPrete and Gangl, 2004). I will propose a set of scenarios that will allow one to directly manipulate the source of the problem: the degree of unobserved heterogeneity. This way one can compare how the results would change if there is a small, moderate, or large amount of unobserved heterogeneity.

This chapter will start with a more detailed discussion of how unobserved heterogeneity can cause bias in the estimates of the effect of the observed variables, even if the unobserved variables are initially non-confounding variables. I will then propose a sensitivity analysis, by specifying a series of scenarios concerning the unobserved variables. The estimation of the effects within these scenarios will be discussed next. Finally, the method will be illustrated by replicating an analysis of the effect of parental background on educational attainment in the Netherlands by De Graaf and Ganzeboom (1993) and in Chapter 2, and assessing how robust their results are for changes in assumptions about unobserved heterogeneity.

7.2 The sequential logit model and two effects of unobserved heterogeneity

The effect of unobserved heterogeneity in a sequential logit model is best explained using an example. Figure 7.1 shows a hypothetical process, which is to be described using a sequential logit model. There are three levels in this process: A, B and C. This process consists of two transitions: the first transition is a choice between A on the one hand and B and C on the other. The second transition is a choice between B and C for those who have chosen B and C in first transition. Figure 7.1 could be a representation of both the educational attainment example and the retirement example in the introduction. In the former case, A would correspond to primary education, B would correspond to secondary education, and C would correspond to tertiary education. In the latter case, A would correspond to retire before age 62, B would correspond to retire between age 62 and 64, and C would correspond to retire after age 64.

Figure 7.1: Hypothetical process



The sequential logit model models the probabilities of passing these transitions. This is done by estimating a logistic regression for each transition on the sub-sample that is at risk, as in equations (7.1) and (7.2). Equation (7.1) shows that the probability labelled p_1 in Figure 7.1 is related to two explanatory variables x and z through the function $\Lambda()$, while equation (7.2) shows the same for the probability labelled p_2 in Figure 7.1. The function $\Lambda()$ is defined such that $\Lambda(u) = \frac{\exp(u)}{1+\exp(u)}$. This function ensures that the predicted probability always remains between 0 and 1, by modelling the effects of the explanatory variables as S-shaped curves. The coefficients of x and z (β_{11}, β_{21}, β_{12}, and β_{22}) can be interpreted as log odds ratios, while the constants (β_{01} and β_{02}) represent the baseline log odds of passing the first and second transitions.

$$p_1 = \Pr(y \in \{B, C\}|x, z) = \Lambda(\beta_{01} + \beta_{11}x + \beta_{21}z) \quad (7.1)$$

$$p_2 = \Pr(y \in \{C\}|x, z, y \in \{B, C\}) = \Lambda(\beta_{02} + \beta_{12}x + \beta_{22}z) \quad (7.2)$$

Table 7.1 turns Figure 7.1 and equations (7.1) and (7.2) into a numerical example. Panel (a) shows the counts, the probabilities of passing, the odds and log odds ratios when z is observed, while panel (b) shows what happens in this example when z is not observed. Both x and z are dichotomous (where low is coded as 0 and high as 1), and during the first transition x and z are independent, meaning that z is not a confounding variable at the first transition. The sequential logit model underlying this example is presented in equations (7.3) and (7.4).

$$\Pr(y \in \{B, C\}|x, z) = \Lambda[\log(.333) + \log(3)x + \log(3)z] \quad (7.3)$$

$$\Pr(y \in \{C\}|x, z, y \in \{B, C\}) = \Lambda[\log(.333) + \log(3)x + \log(3)z] \quad (7.4)$$

Consider the first transition in panel (a). The constant in the logistic regression equation is the log odds of passing for the group with value 0 for all explanatory vari-

Table 7.1: Example illustrating the consequences of not observing a non-confounding variable (z)

(a) while observing z

transition	z	x	y: A	y: B, C	N	Pr(pass)	odds(pass)	log odds ratio
1	low	low	300	100	400	0.25	0.333	log(3)
		high	200	200	400	0.5	1	
	high	low	200	200	400	0.5	1	log(3)
		high	100	300	400	0.75	3	
			B	C				
2	low	low	75	25	100	0.25	0.333	log(3)
		high	100	100	200	0.5	1	
	high	low	100	100	200	0.5	1	log(3)
		high	75	225	300	0.75	3	

(b) without observing z

transition	x	y: A	y: B, C	N	Pr	odds	log odds ratio
1	low	500	300	800	0.375	0.6	log(2.778)
	high	300	500	800	0.625	1.667	
		B	C				
2	low	175	125	300	0.417	0.714	log(2.6)
	high	175	325	500	0.65	1.857	

ables, so the constant is in this case log(.333). The effect of x in a logistic regression equation is the log odds ratio. Within the low z group, the odds of passing for the low x group is .333 and the odds of passing for the high x group is 1, so odds ratio is $\frac{1}{.333} = 3$, and the log odds ratio is log(3). The effect of x in the high z group is also log(3), so there is no interaction effect between x and z. The effect of z can be calculated by comparing the odds of passing for a high z and a low z individual within the low x group, which results in a log odds ratio of log(3). There is no interaction between x and z, so the log odds ratio for z within the high x group is also log(3). Panel (b) shows what happens if one only observes x and y but not z. For example, in that case 300 + 200 = 500 low x persons are observed to have failed the first transition and 100 + 200 = 300 low x persons are observed to have passed the first transition. The resulting counts are used to calculate the probabilities, odds, and log odds ratios. Panel (b) shows that the log odds ratios of x are smaller than those computed in panel (a). Leaving z out of the model thus resulted in an underestimation of the effect of x for both the first and the second transition, even though z was initially uncorrelated with x.

This example can be used to illustrate both mechanisms through which unobserved heterogeneity can lead to biased estimates of the individual-level effects. First, the selection mechanism can explain part of the underestimation of the effect of x at the second transition. A characteristic of the sequential logit model is that even if z is not a confounding variable during the first transition, it will become a confounding variable during the later transitions (Mare, 1980; Cameron and Heckman, 1998). The example was created such that z and x are independent during the first transition, as the distribution of z is equal for both the low x group and high x group. As a consequence, z cannot be a confounding variable during the first transition. But this is no longer true during the second transition. For the high x group, the proportion of persons with a high z is $300/500 = .6$, while for the low x group that proportion is $200/300 = .667$. The selection at the first transition has thus introduced a negative correlation between x and z, and z has become a confounding variable. If one does not observe z, and thus can not control for z, one would expect to underestimate the effect of x at the second transition. This could in part explain the underestimation of the effect of x in the second transition in panel (b) of Table 7.1, but not the underestimation of the effect of x in the first transition.

The averaging mechanism can explain the underestimation of the effect of x during the first transition. The models implicit in panels (a) and (b) have subtly different dependent variables: in panel (a) one is modelling the probability that an *individual* passes the transitions, while in panel (b) one models the *average* probability of passing the transitions. The two result in different estimates because the relationship between the unobserved variables and the probabilities is non-linear. This issue is discussed in terms of the sequential logit model by Cameron and Heckman (1998). It also occurs in other non-linear models, and has been discussed by Neuhaus et al. (1991), Allison (1999) and Mood (2010). It is also closely related to the distinction between population average or marginal models on the one hand and mixed effects or subject specific models on the other (Fitzmaurice et al. 2004, chapter 13; Agresti 2002, chapter 12). The averaging of the probabilities can be seen in Table 7.1: for example the probability of passing transition 2 for low x individuals when not controlling for z is (100 $\times$ 0.25 + 200 $\times$ 0.5)/300 = 0.417. The consequence of this is that if we think that equations (7.1) and (7.2) form the true model for the probabilities of passing the transitions, then the true model for the probabilities averaged over z should be represented by equations (7.5) and (7.6), where $\mathrm{E}_z(u)$ is the average of u over z. Instead, the model represented by equations (7.7) and (7.8) are estimated when z is not observed and z is thus left out of the model. The two models are not the equivalent because $\Lambda()$ is a non-linear transformation. Neuhaus and Jewell (1993) give an approximation of how β_{11}^* and β_{12}^* deviate from β_{11} and β_{12}: β_{11}^* and β_{12}^* will be smaller than β_{11} and β_{12}, and the difference between the estimates β_{11}^* and β_{12}^* and

the estimates β_{11} and β_{12} will increase when the variances of $\beta_{21}z$ and $\beta_{22}z$ increase and when the probability of passing is closer to 50%.

$$\begin{aligned} \mathrm{E}_z(\Pr[y \in \{B,C\}|x,z]) &= \mathrm{E}_z(\Lambda(\beta_{01} + \beta_{11}x + \beta_{21}z)) \quad (7.5) \\ \mathrm{E}_z(\Pr[y \in \{C\}|x,z,y \in \{B,C\}]) &= \mathrm{E}_z(\Lambda(\beta_{03} + \beta_{12}x + \beta_{22}z)) \quad (7.6) \end{aligned}$$

$$\begin{aligned} \mathrm{E}_z(\Pr[y \in \{B,C\}|x,z]) &= \Lambda(\beta^*_{01} + \beta^*_{11}x) \quad (7.7) \\ \mathrm{E}_z(\Pr[y \in \{C\}|x,z,y \in \{B,C\}]) &= \Lambda(\beta^*_{02} + \beta^*_{12}x) \quad (7.8) \end{aligned}$$

7.3 A sensitivity analysis

The previous section discussed what kind of problems unobserved variables might cause. The difficulty with finding a solution for these problems is that it is obviously challenging to control for something that has not been observed. One possible solution is to perform a sensitivity analysis: specify a number of plausible scenarios concerning the unobserved variables, and estimate the effects within each scenario. The aim of this type of analysis is not to get an empirical estimate of the effect *per se*, but to assess how important assumptions are for the estimated effect and to get a feel for the range of plausible values for the effect. There are many potential problems that could all simultaneously influence the results of an analysis and whose influence could all be investigated using sensitivity analysis. However, to give the analysis focus it is often better to narrow down the scope of the sensitivity analysis by concentrating on a specific subset of potential problems. For example, the aim of the sensitivity analysis proposed in this chapter is to assess the sensitivity to the effect of unobserved heterogeneity through the selection mechanism and averaging mechanism.

A key step in creating such scenarios is to create a set of reasonable scenarios concerning the unobserved variable z. In the example in the previous section, z was assumed to be dichotomous, because that would result in an easy numerical example. When creating the scenarios, it is more useful to think about z as not being a single unobserved variable but as a (weighted) sum of all the unobserved variables. Such a sum of random variables can usually be well approximated by a normal distribution, even if the constituent variables are non-normally distributed. So, it is reasonable to represent the distribution of the composite unobserved variable with a normal distribution. There are two equivalent ways of thinking about the scale of this compound unobserved variable. It is sometimes convenient to think of the resulting variable as

being standardized, such that mean is 0 and the standard deviation is 1. This way the 'effect' — call that γ — can be compared with the effects of standardized observed variables to get a feel for the range of reasonable values of this 'effect'. Alternatively, it is possible to think of the composite unobserved variable as just being an unstandardized random variable or error term. In this case, the standard deviation of this random variable is the same as γ. The standardized unobserved variable will be referred to as z, while the unstandardized unobserved variable will be referred to as ε in order to distinguish between the two. The two are related in the following way: $\gamma z = \varepsilon$.

In this chapter I will propose a set of scenarios based on this representation of the unobserved variable. This basic scenario is introduced in equations (7.9) till (7.12). In this example there are two transitions, with the probabilities of passing these transitions influenced by two variables x and z, where z is as defined above. The observed dependent variables are the probabilities of passing the two transitions averaged over z. So by estimating models (7.9) and (7.11), one can recover the true effects of x. To estimate it, all one needs to know is the distribution of γz $(= \varepsilon)$ and to integrate over this distribution, as in equations (7.10) and (7.12). The mean of ε will be set at 0 and a standard deviation equal to γ, which is *a priori* fixed in the scenario. Furthermore, it assumes that a person's value on ε will not change over the transitions, implicitly assuming that both the value on z and the effect of z (γ) will not change over the transitions[2].

$$\begin{aligned}
\mathrm{E}_\varepsilon(\Pr[y \in \{B,C\}|x,\varepsilon]) &= \mathrm{E}_\varepsilon(\Lambda(\beta_{01} + \beta_{11}x + \underbrace{\gamma z}_{\varepsilon})) && (7.9)\\
&= \int \Lambda(\beta_{01} + \beta_{11}x + \varepsilon) f(\varepsilon) d\varepsilon && (7.10)\\
\mathrm{E}_\varepsilon(\Pr[y \in \{C\}|x,\varepsilon, y \in \{B,C\}]) &= \mathrm{E}_\varepsilon(\Lambda(\beta_{02} + \beta_{12}x + \underbrace{\gamma z}_{\varepsilon})) && (7.11)\\
&= \int \Lambda(\beta_{02} + \beta_{12}x + \varepsilon) \\
&\quad f(\varepsilon|y \in \{B,C\}) d\varepsilon && (7.12)
\end{aligned}$$

The effects in each scenario are estimated using maximum likelihood. Referring back to Figure 7.1, the likelihood function for an individual i can be written as equa-

[2] All these assumptions can be relaxed, but relaxing these assumptions will quickly lead to an unmanageable number of scenarios. Moreover, these complications would not contribute to the aim of these scenarios, which assess the sensitivity of estimates to unobserved heterogeneity through the selection mechanism and averaging mechanism.

tion (7.13), that is, the probability of observing someone with value A equals the probability of failing the first transition, the probability of observing someone with value B equals the probability of passing the first transition and failing the second transition, and the probability of observing someone with value C equals the probability of passing both transitions.

$$L_i = \begin{cases} 1 - p_{1i} & \text{if } y_i = A \\ p_{1i} \times (1 - p_{2i}) & \text{if } y_i = B \\ p_{1i} \times p_{2i} & \text{if } y_i = C \end{cases} \tag{7.13}$$

By replacing p_{1i} with equation (7.10) and p_{2i} with equation (7.12), one gets a function that gives the probability of an observation, given the parameters β. This probability can be computed for each observation and the product of these form the probability of observing the data, given a set of parameters. Maximizing this function with respect to the parameters give the maximum likelihood estimates. These estimates include the true effects of the variable of interest x assuming that the model for the unobserved heterogeneity, in particular the standard deviation of ε, is correct.

The difficulty with this likelihood is that there are no closed form solutions for the integrals in equations (7.10) and (7.12). This can be resolved by numerically approximating these integrals using maximum simulated likelihood (Train, 2003). Maximum simulated likelihood uses the fact that the integral is only there to compute a mean probability. This mean can be approximated by drawing at random many values for ε from the distribution of ε, computing the probability of passing a transition assuming that this randomly drawn value is the true value of ε, and then computing the average of these probabilities. This approach can be further refined by realizing that using true random draws is somewhat inefficient as these tend rather to cluster. Increasing the efficiency is important as these integrals need to be computed for each observation, meaning that these simulations need to be repeated for each observation. One can cover the entire distribution with less draws if one can use a more regular sequence of numbers. An example of a more regular sequence of numbers is a Halton (1960) sequence. A Halton sequence will result in a more regular series of quasi-random draws from a uniform distribution. These quasi-random draws can be transformed into quasi-draws from a normal distribution by applying the inverse cumulative normal distribution function. These are then used to compute the average probability of passing the first transition, as is shown in equation (7.14), where m represents the number of draws from the distribution of ε. At subsequent transitions, the distribution of ε is no longer a normal distribution, but conditional on being at risk. The integral over this distribution is computed by drawing ε from a normal distribution as before, but then computing a weighted mean whereby each draw is given a weight equal to

the probability of being at risk assuming that that draw was the true ε. In the appendix to this chapter I show that this is a special case of importance sampling (Robert and Casella, 2004, 90–107). This procedure is implemented in the `seqlogit` package (Buis, 2007b) in Stata (StataCorp, 2007), using the facilities for generating Halton sequences discussed by Drukker and Gates (2006). This package is documented in Technical Materials II.

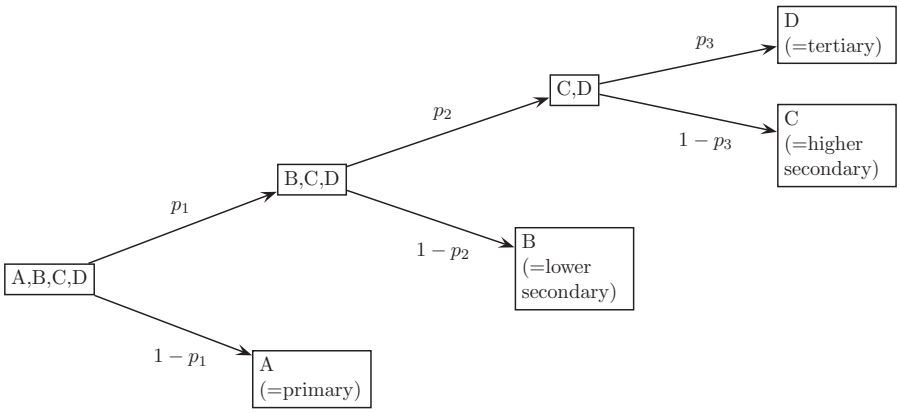
$$\mathrm{E}_{\varepsilon}(\Pr(y \in \{B,C\}|x,\varepsilon)) \approx \frac{1}{m}\sum_{j=1}^{m}\Lambda(\beta_{01}+\beta_{11}x+\varepsilon_j) \tag{7.14}$$

$$\mathrm{E}_{\varepsilon}(\Pr(y \in \{C\}|x,\varepsilon,y \in \{B,C\}) \approx \frac{\sum_{j=1}^{m}\left[\Pr(y\in\{B,C\}|x,\varepsilon_j)\Lambda(\beta_{02}+\beta_{12}x+\varepsilon_j)\right]}{\sum_{j=1}^{m}\Pr(y\in\{B,C\}|x,\varepsilon_j)} \tag{7.15}$$

7.4 An example: The effect of family background on educational attainment in the Netherlands

An important application for the sequential logit model is the study of the influence of family background on educational attainment (for recent reviews see: Breen and Jonsson, 2005; Hout and DiPrete, 2006). The potential problems that unobserved variables can cause were recognized from the time that the sequential logit model was introduced in this literature (Mare, 1979, 1980, 1981), but interest in this issue has been revived by the critique from Cameron and Heckman (1998). However, only a limited number of empirical studies have tried to actually account for unobserved heterogeneity (for exceptions see: Mare, 1993; Rijken, 1999; Chevalier and Lanot, 2002; Lauer, 2003; Arends-Kuenning and Duryea, 2006; Colding, 2006; Lucas et al., 2007; Holm and Jæger, 2008). The method proposed in this paper will be illustrated by replicating an analysis that does not control for unobserved heterogeneity by De Graaf and Ganzeboom (1993) and in Chapter 2 of the effect of father's occupational status and education on transition probabilities between educational levels in the Netherlands, and assessing how sensitive the conclusions are to assumptions about unobserved heterogeneity. The original study by De Graaf and Ganzeboom (1993) was part of an influential international comparison of the effect of family background on educational attainment (Shavit and Blossfeld, 1993). It used 10 Dutch surveys that were post-harmonized as part of the International Stratification and Mobility File [ISMF] (Ganzeboom and Treiman, 2009). In Chapter 2 I updated this analysis by using an additional 33 Dutch surveys that have since been added to the ISMF.

Figure 7.2: Simplified model of the Dutch education system



7.4.1 The data

The total of 43 surveys were held between 1958 and 2006. Only male respondents older than 25 are used in the analysis. These surveys contain 35,846 men with valid information on all the variables used in the model. Family background is measured as the father's occupational status and the father's highest achieved level of education. Time was measured by 10-year birth cohorts covering the cohorts that were born between 1891–1980. The main effect of time is added as a set of dummies, while the effects of the family background variables is allowed to change linearly over the cohorts.

The father's occupational status was measured using the International Socio-Economic Index (ISEI) of occupational status (Ganzeboom and Treiman, 2003), which originally ranged between 10 and 90 and was recoded to range between 0 and 8. In concordance with De Graaf and Ganzeboom (1993) and Chapter 2, education of both the father and the respondent were measured in four categories: primary education (LO), lower second secondary education (LBO and MAVO), higher secondary education (HAVO, MBO, and VWO), and tertiary education (HBO and WO). The value of the father's highest achieved level of education was created by giving these educational categories the numerical values 1 till 4. The transitions that were studied by De Graaf and Ganzeboom (1993) and in Chapter 2 are: 1) from primary education or less to a diploma in secondary or tertiary education; 2) from a diploma in lower secondary education to a diploma in higher secondary or tertiary education; 3) from a diploma in higher secondary education to completed tertiary education. These transitions are displayed in Figure 7.2.

7.4.2 The results

The effects of father's occupational status and education are estimated for four scenarios, and the results are represented in the different columns in Table 7.2. The first scenario assumes that the standard deviation of ε is zero, which is a replication of the model used by De Graaf and Ganzeboom (1993) and in Chapter 2. This replication shows three main patterns. First, both father's occupational status and father's education have a positive effect on the probability of passing transitions. Second, this effect decreases over transitions. Third, the effect of father's education decreases over cohorts during all three transitions while the effect of father's occupational status clearly decreases over cohorts for the first transition, but the trend is non-significant negative during the second transition and non-significant positive during the third transition. These patterns are the same as those found by De Graaf and Ganzeboom (1993) and in Chapter 2 with the exception of the significant negative trend in the effect of father's education during the third transition, which was not found to be significant by De Graaf and Ganzeboom (1993).

The remaining three scenarios assume that the standard deviation of ε is .5, 1, and 2. As was discussed before, the standard deviations represent the effects (log odds ratios) if the unobserved variable z is a standardized variable. To put these scenarios into perspective, one can look at the effects of father's occupational status and education when both are standardized in the earliest cohort at the first transition, when the effects are largest. These standardized effects are .823 for father's occupational status and 1.453 for father's education[3]. So, the values .5, 1, and 2 capture a reasonable range of values for the effect of a standardized unobserved variable.

The results from the different scenarios, as presented in the remaining columns of Table 7.2, show that the qualitative conclusions remain unchanged, that is, those effects that were significant remained significant and those that were not significant remained not significant. However, the size of the effects of father's occupational status and education and their trends did change over the scenarios: the effects increased as the amount of unobserved heterogeneity increased, while the trends in the effects over time became more negative, and the decrease in the effects over transitions becomes less pronounced. This is also shown in Figures 7.3 and 7.4. In addition, these figures show that difference between the scenarios decreased over time, indicating that the bias due to unobserved heterogeneity decreased over time. This is particularly strong for the first transition.

In section 7.2 I discussed that unobserved heterogeneity could influence the results through two mechanisms. First, the averaging mechanism is based on the fact

[3]The effects of the unstandardized variables are presented in Table 7.2, and the standard deviation of father's occupational status is 1.55 and the standard deviation of father's education is 1.01.

Table 7.2: Log odds ratios in models for men assuming different degrees of unobserved heterogeneity (the main effects of the cohort dummies and the constant are not displayed)

	sd(ε) = 0	sd(ε) = .5	sd(ε) = 1	sd(ε) = 2
primary v lower secondary				
father's education	1.439	1.496	1.641	2.092
	(11.50)	(11.56)	(11.70)	(12.10)
father's education X cohort	-0.117	-0.124	-0.142	-0.192
	(-4.80)	(-4.96)	(-5.28)	(-5.87)
father's occupation	0.531	0.558	0.628	0.833
	(13.08)	(13.22)	(13.46)	(13.73)
father's occupation X cohort	-0.057	-0.061	-0.070	-0.097
	(-6.34)	(-6.57)	(-7.02)	(-7.60)
lower secondary v higher secondary				
father's education	0.713	0.796	0.995	1.512
	(11.79)	(12.50)	(13.86)	(15.94)
father's education X cohort	-0.026	-0.034	-0.051	-0.092
	(-2.34)	(-2.88)	(-3.88)	(-5.31)
father's occupation	0.294	0.333	0.424	0.655
	(8.10)	(8.73)	(9.96)	(11.90)
father's occupation X cohort	-0.010	-0.014	-0.023	-0.045
	(-1.49)	(-2.01)	(-2.98)	(-4.42)
higher secondary v tertiary				
father's education	0.445	0.539	0.748	1.252
	(7.11)	(8.14)	(10.05)	(12.99)
father's education X cohort	-0.031	-0.039	-0.057	-0.099
	(-2.78)	(-3.34)	(-4.31)	(-5.70)
father's occupation	0.149	0.187	0.275	0.486
	(3.41)	(4.07)	(5.34)	(7.42)
father's occupation X cohort	0.010	0.007	0.001	-0.011
	(1.25)	(0.87)	(0.15)	(-0.97)

(z-values in parentheses)

Figure 7.3: The effect of father's occupational status

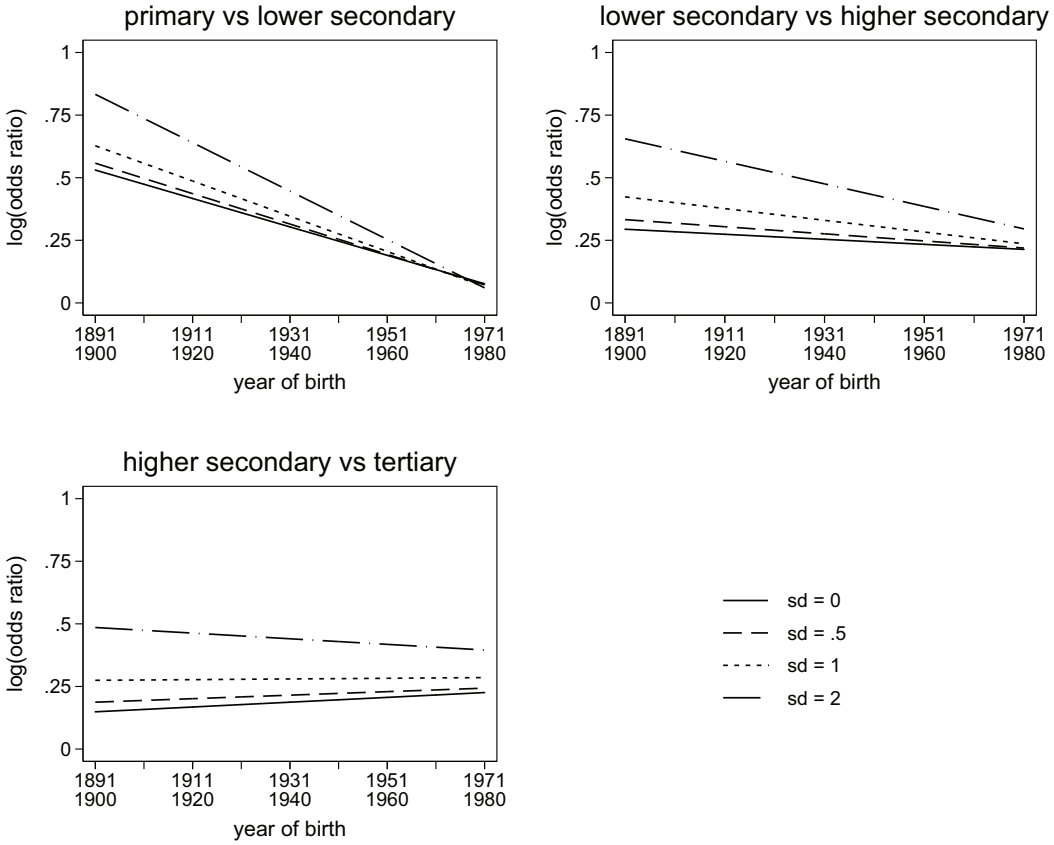
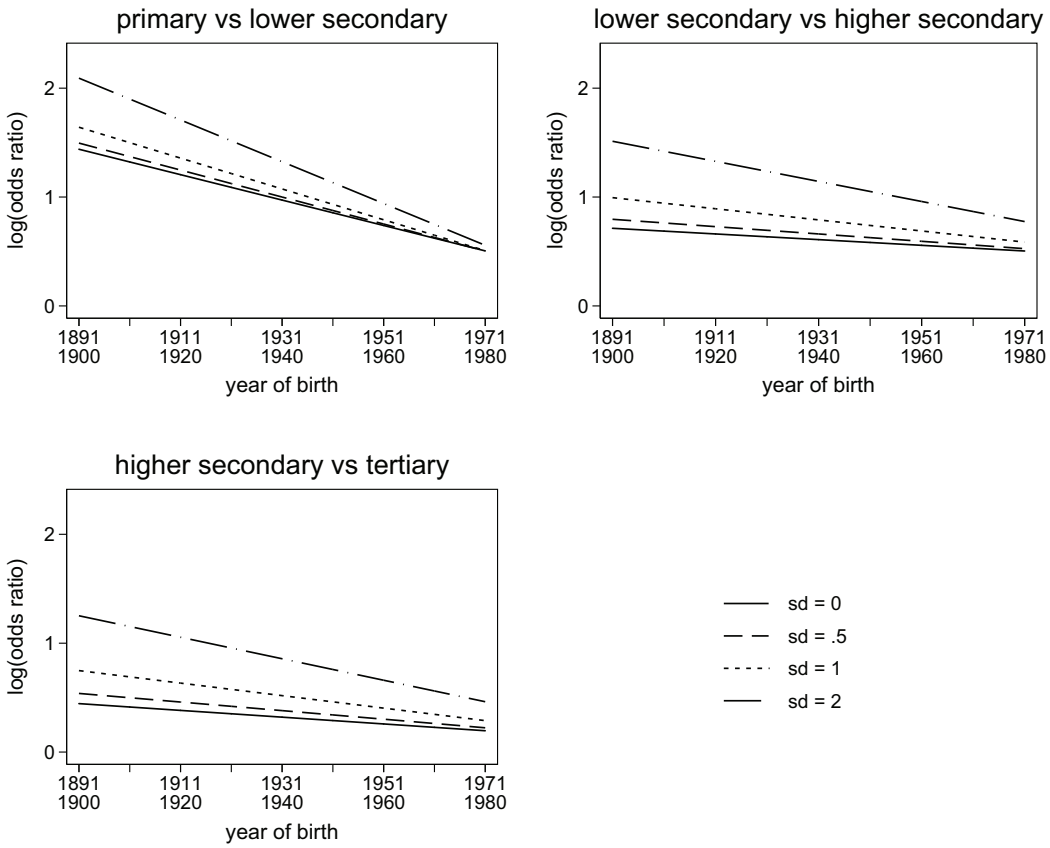


Figure 7.4: The effect of father's education



that a model that leaves out the unobserved variables models the average probability of passing the transitions rather than an individual's probability of passing. This will lead to an underestimation of the effect if one leaves the variable out of the model, and this bias will be larger when the variance of the unobserved variable increases and when the probability of passing is closer to 50% (Neuhaus and Jewell, 1993). Second, the selection mechanism is based on the fact that after the first transition the unobserved variable becomes correlated with the observed variables. This means that at later transitions, leaving the unobserved variable out of the model will result in omitted variable bias, even if the unobserved variable was not a confounding variable at the first transition. A key element in both mechanisms is the distribution of the unobserved variable. Table 7.3 shows how the distribution of the unobserved variable changes over the transitions for the different scenarios for men born between 1931 and 1940 (the largest cohort in the data). The first row shows the proportion of respondents at risk of passing this transition, which indicates how selective a transition is. The second and third set of rows shows for each scenario and transition the correlation between the unobserved variable and father's occupational status, and between the unobserved variable and father's education, respectively. This correlation captures the selection mechanism. At the first transition this correlation is by definition 0, but at later transitions it becomes negative, leading to an underestimation of the effect of father's occupational status and education at the later transitions. The correlation becomes larger at later transitions and when the variance of the unobserved variable increases. The correlation between ε and father's education is stronger than the correlation between ε and the father's occupational status. The reason for this is that the correlation is the result of the selection on all the variables at the earlier transitions, and the selection on father's education is stronger than the selection on father's occupational status (the standardized coefficients for the main effects are, as was mentioned before, 1.453 and .823 respectively). The fourth set of rows shows that the variance of the unobserved variable, which plays a key role in the averaging mechanism, and which decreases somewhat over the transitions, but not much. The fifth set of rows shows that the respondents score higher than average on the unobserved variable at the higher transition.

Table 7.3: Changes in the distribution of the unobserved variable over the transitions for men born between 1931 and 1940

		primary v lower secondary	lower secondary v higher secondary	higher secondary v tertiary
Pr(at risk)		1	.837	.487
corr(ε, father's occupation)	sd(ε) = 0	0	0	0
	sd(ε) = 0.5	0	-0.028	-0.070
	sd(ε) = 1	0	-0.051	-0.124
	sd(ε) = 2	0	-0.081	-0.187
corr(ε, father's education)	sd(ε) = 0	0	0	0
	sd(ε) = 0.5	0	-0.048	-0.111
	sd(ε) = 1	0	-0.087	-0.193
	sd(ε) = 2	0	-0.134	-0.282
sd(ε)	sd(ε) = 0	0	0	0
	sd(ε) = 0.5	0.5	0.492	0.480
	sd(ε) = 1	1	0.950	0.883
	sd(ε) = 2	2	1.764	1.531
mean(ε)	sd(ε) = 0	0	0	0
	sd(ε) = 0.5	0	0.038	0.132
	sd(ε) = 1	0	0.143	0.460
	sd(ε) = 2	0	0.460	1.313

Table 7.3 gives an idea of the distribution of the unobserved variable at one point in time, but it cannot explain why this bias changed over time, as was shown in Figures 7.3 and 7.4. The way unobserved heterogeneity influences the results is a function of the proportion of respondents that are at risk at each transition and these have changed considerably over time as is shown in Figure 7.5. As in most other countries, younger cohorts will on average receive more education than the older cohorts, so the proportion of respondents at risk increases over time. Figure 7.5 also explains why the bias in the first transition decreases. The bias in the first transition is due to the averaging mechanism, and the bias due to the averaging mechanism will decrease when the probability of passing approaches 1 (or 0) (Neuhaus and Jewell, 1993). The proportion of respondents that passed the first transition is the proportion at risk of passing the second transition. Figure 7.5 shows that this proportion increased dramatically and is now virtually 1, thus leading to a reduction in the size of the bias. Figure 7.6 showed how the correlation between the father's occupational status and education and the unobserved variable changed over time. It shows that this correlation strongly decreased over time as the higher transitions became less selective, and thus that the bias due to the selection mechanism decreased over time. Figure 7.7 shows that the standard deviation of the unobserved variable hardly changes over time. Figure 7.8 shows how the mean of the unobserved variable decreases at each subsequent transition and how these transitions have become less selective over time.

In summary, this replication showed that the qualitative conclusions from De Graaf and Ganzeboom (1993) and Chapter 2 are largely robust against assumptions on unobserved heterogeneity. However, the scenarios also showed that the size of the effects and the trends are likely to have been underestimated because the original sequential logit models estimated the effect on the average probability of passing rather than on an individual's probability of passing, and because the unobserved variable and the observed variables became negatively correlated at the higher transitions.

Figure 7.5: The proportion of respondents at risk of passing each transition

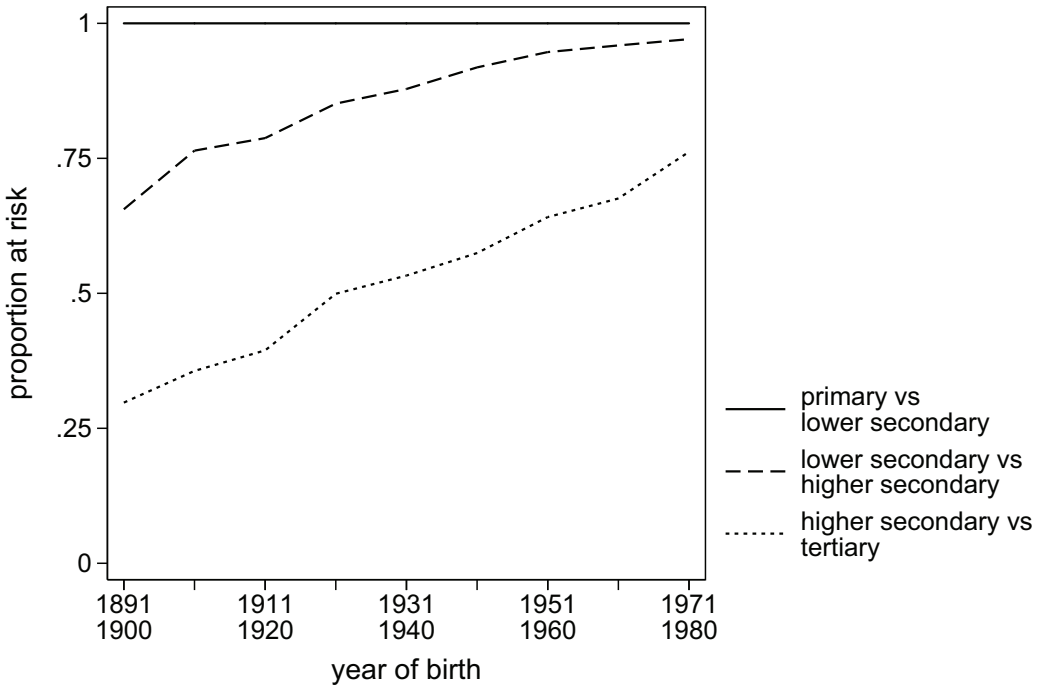


Figure 7.6: The correlation between the unobserved variable and father's occupational status and father's education

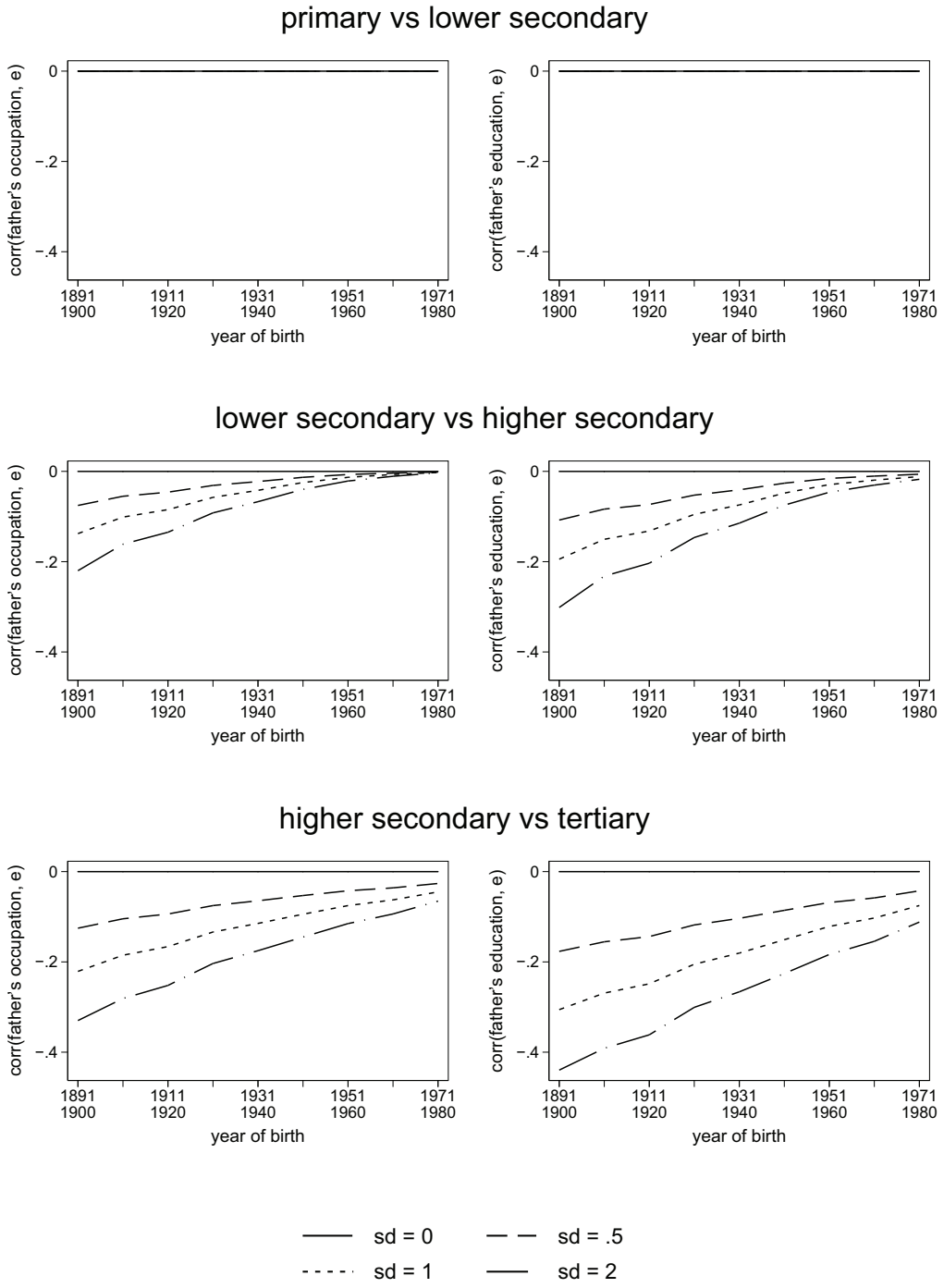


Figure 7.7: The standard deviation of the unobserved variable

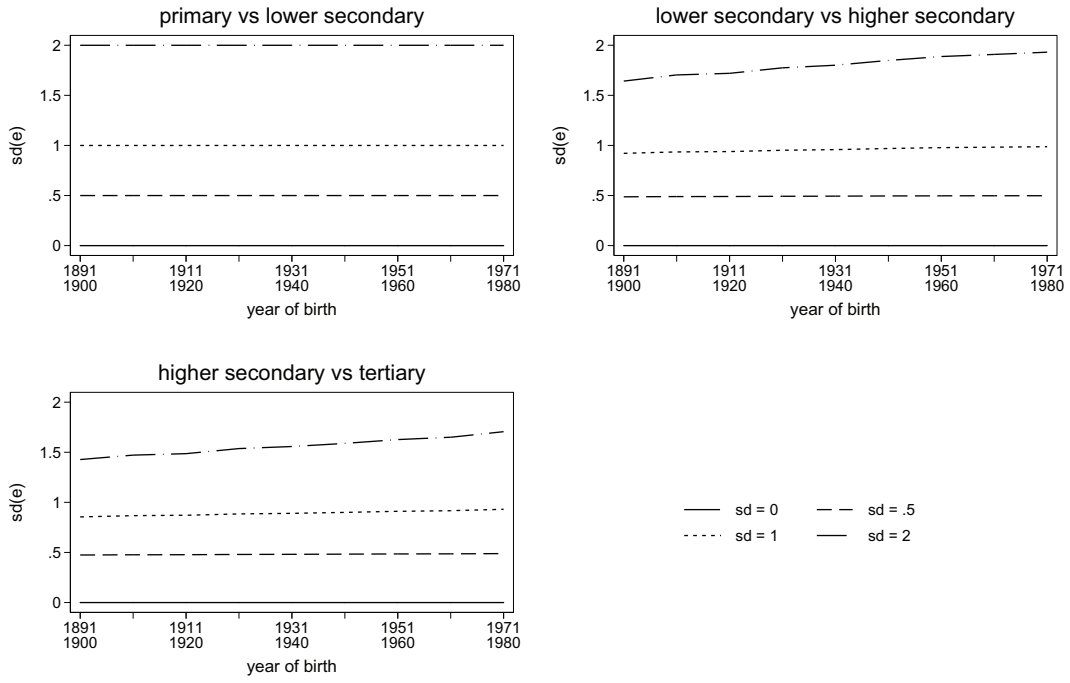
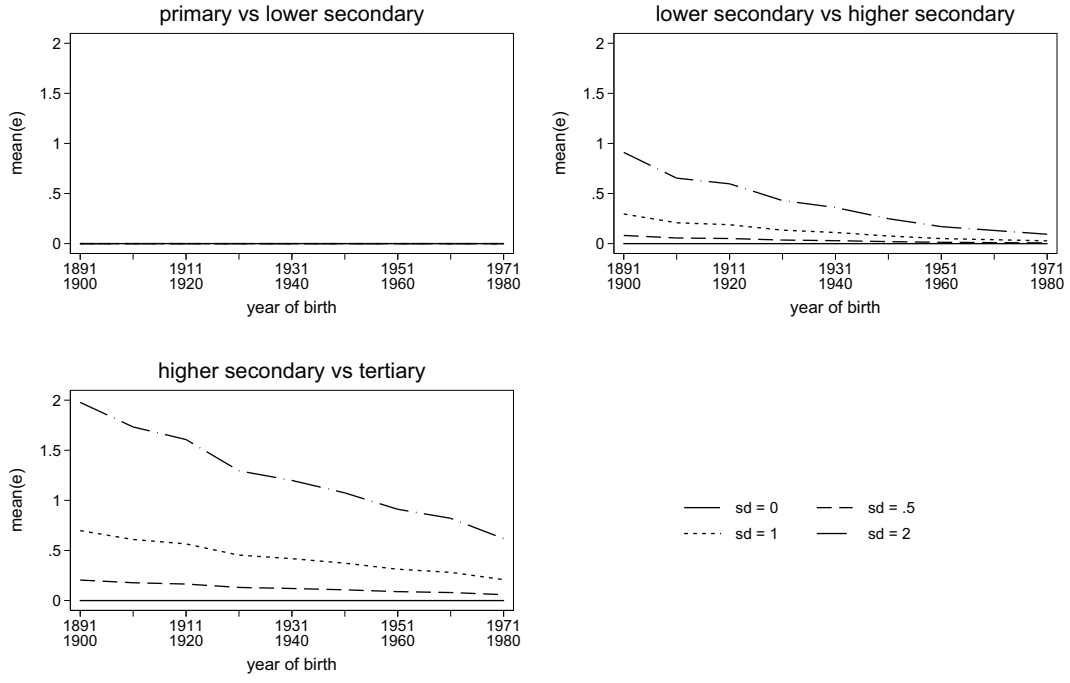


Figure 7.8: The mean of the unobserved variable



7.5 Conclusion and discussion

The aim of this chapter is to present a sensitivity analysis that can be used to investigate the consequences of unobserved variables in a sequential logit model, and in particular the consequences of leaving a non-confounding variable out of a sequential logit model as discussed by Cameron and Heckman (1998). The bias that these unobserved variables cause are shown to be the result of two mechanisms: first, the averaging mechanism is based on the fact that when a variable is left out of the model, one models the probability of passing the transitions averaged over the variable that is left out. As a consequence, just leaving the unobserved variable out of the model will lead to estimates of effects of the observed variables on the probability of passing the transitions averaged over the unobserved variables rather than the effects on the individual's probability of passing. These two are different because the unobserved variable is related to the probabilities through a non-linear function. Second, the selection mechanism is based on the fact that a variable that is not a confounding variable at the first transition is likely to become a confounding variable at the later transitions. The reason for this is that the process of selection at the earlier transitions will introduce correlation between the observed and unobserved variables.

The method proposed in this chapter to investigate the consequences of unobserved heterogeneity is to perform a sensitivity analysis by specifying a set of scenarios regarding the extent of unobserved heterogeneity, and estimating the effects of the observed variables given those scenarios. This will not give an empirical estimate of the effects of interest, but does give an idea about the sensitivity of the estimates to assumptions about unobserved heterogeneity, and direction of the bias, the size of the bias, and the range of likely values of the effect. The scenarios that have been proposed in this chapter are constructed in the following way: the unobserved variable is normally distributed, for each individual the value of this unobserved variable is assumed to remain constant over the transitions, and the effect of the unobserved variable also remains constant over the transitions. The scenarios differ from one another with respect to the variance of the unobserved variable. This way one can compare what happens to the effects of the observed variables when there is a small, medium, and large amount of unobserved heterogeneity. Moreover, it is possible to recover the distribution of the unobserved variable at the later transitions. This makes it possible to see how, in each scenario, the correlation between the observed and unobserved variables change over the transitions, and/or over a third variable, for example time. The effects of the observed variables within each scenario are estimated by maximum likelihood. The likelihood is defined by integrating over the unobserved variable, which is done using Maximum Simulated Likelihood (Train, 2003).

This method was illustrated by replicating a study by De Graaf and Ganzeboom

(1993) and in Chapter 2 on the effect of the father's occupational status and education on the offspring's educational attainment. The proposed analysis showed that the results of statistical tests were rather robust to changes in the assumptions about unobserved heterogeneity, but that the effects of both the father's occupational status and the father's education were likely to be underestimated, as these effects were stronger in scenarios with more unobserved heterogeneity. Scenarios with more unobserved heterogeneity also resulted in a stronger downward trend over time in the effect of father's occupational status and education. The decrease in the effect of father's occupational status and education over transitions became less in scenarios with more unobserved heterogeneity. This indicates that the commonly found pattern of decreasing effects of family background variables over transitions is at least in part due to unobserved heterogeneity.

This chapter can be seen as part of a larger effort aimed at obtaining an empirical estimate of the causal effect of family background while controlling for unobserved variation between individuals (Mare, 1993, 1994; Cameron and Heckman, 1998; Lucas et al., 2007; Holm and Jæger, 2008). The challenge of this literature is that it tries to solve an unsolvable problem, since obtaining an empirical estimate is by definition incompatible with controlling for unobserved variation. On the one hand this means that it is very unlikely that a single study can build a completely convincing empirical argument for such an effect. On the other hand, that does not mean that estimates obtained in these studies contain no information whatsoever. The key is that each of these methods exploits different parts of the data to get an approximation of the effect. For example, Mare (1993, 1994) uses the nesting of individuals within families, Lucas et al. (2007) and Holm and Jæger (2008) use the presence of instrumental variables, and Mare (2006) uses the strong assumption that all changes in the effect of the explanatory variables over transitions is due to unobserved heterogeneity. In the long run, these differences in strategy can be used to get a plausible range for the causal effect of family background by collecting a sufficient body of evidence using these different methods, followed by an analysis of how the differences in strategy has led to the differences and similarities in the conclusions of these studies.

Appendix: Sampling from the distribution of ε conditional on having passed the previous transitions

One method of sampling from a distribution is importance sampling (Robert and Casella, 2004, 90–107). This appendix will show that the method used in this chapter is a special case of importance sampling. The idea behind importance sampling is that instead of sampling from the distribution of interest $f(\varepsilon)$ one draws samples from another distribution $g(\varepsilon)$, and compute the mean by weighting each draw by $\frac{f(\varepsilon_j)}{g(\varepsilon_j)}$, so one could approximate $\mathrm{E}_\varepsilon[\Lambda(\beta_{02}+\beta_{12}x+\varepsilon)]$ with equation (7.16).

$$\mathrm{E}_\varepsilon[\Lambda(\beta_{02}+\beta_{12}x+\varepsilon)] \approx \frac{1}{m}\sum_{j=1}^{m}\frac{f(\varepsilon_j)}{g(\varepsilon_j)}\Lambda(\beta_{02}+\beta_{12}x+\varepsilon) \tag{7.16}$$

In this chapter the distribution of interest is the distribution conditional on being at risk, while the other distribution is the distribution not conditional on being at risk. These distributions are independent of x, so the conditioning on x in equation (7.17) is superfluous, but this will prove useful later on.

$$\mathrm{E}_\varepsilon[\Lambda(\beta_{02}+\beta_{12}x+\varepsilon)] \approx \frac{1}{m}\sum_{j=1}^{m}\frac{f(\varepsilon_j|x, y\in\{B,C\})}{f(\varepsilon_j|x)}\Lambda(\beta_{02}+\beta_{12}x+\varepsilon) \tag{7.17}$$

Instead of using equation (7.17) directly, the integral is computed using equation (7.18). The aim of this appendix is to show that these two are equivalent.

$$\mathrm{E}_\varepsilon[\Lambda(\beta_{02}+\beta_{12}x+\varepsilon)] \approx \frac{\sum_{j=1}^{m}\left[\Pr(y\in\{B,C\}|x,\varepsilon_j)\Lambda(\beta_{02}+\beta_{12}x+\varepsilon)\right]}{\sum_{j=1}^{m}\Pr(y\in\{B,C\}|x,\varepsilon_j)} \tag{7.18}$$

The denominator of equation (7.18) can be rewritten as in equation (7.19), which leads to equation (7.20)

$$\begin{aligned}\sum_{j=1}^{m}\Pr(y\in\{B,C\}|x,\varepsilon_j) &= m\frac{\sum_{j=1}^{m}\Pr(y\in\{B,C\}|x,\varepsilon_j)}{m}\\ &\approx m\Pr(y\in\{B,C\}|x)\end{aligned} \tag{7.19}$$

$$\mathrm{E}_{\varepsilon}[\Lambda(\beta_{02}+\beta_{12}x+\varepsilon)] \approx \frac{1}{m}\sum_{j=1}^{m}\frac{\Pr(y\in\{B,C\}|x,\varepsilon_j)}{\Pr(y\in\{B,C\}|x)}\Lambda(\beta_{02}+\beta_{12}x+\varepsilon) \quad (7.20)$$

Comparing equations (7.17) and (7.20) indicates that the problem can be simplified to showing that equation (7.21) is true.

$$\frac{f(\varepsilon_j|x,y\in\{B,C\})}{f(\varepsilon_j|x)} = \frac{\Pr(y\in\{B,C\}|x,\varepsilon_j)}{\Pr(y\in\{B,C\}|x)} \quad (7.21)$$

Equation (7.21) can be rewritten as equation (7.22). Using Bayes' theorem, equation (7.22) can be rewritten as equation (7.23). Equation (7.23) is true, thus showing that equations (7.17) and (7.18) are equivalent. Notice, however, that this is based on the approximation in equation (7.19), which will get better as the number of samples m increases.

$$f(\varepsilon_j|x,y\in\{B,C\})\Pr(y\in\{B,C\}|x) = \Pr(y\in\{B,C\}|x,\varepsilon_j)f(\varepsilon_j|x) \quad (7.22)$$

$$f(\varepsilon_j\cap y\in\{B,C\}|x) = f(y\in\{B,C\}\cap\varepsilon_j|x) \quad (7.23)$$