Paper 184-31

Fixed Effects Regression Methods In SAS®

Paul D. Allison, University of Pennsylvania, Philadelphia, PA

ABSTRACT

Fixed effects regression methods are used to analyze longitudinal data with repeated measures on both independent and dependent variables. They have the attractive feature of controlling for all stable characteristics of the individuals, whether measured or not. This is accomplished by using only within-individual variation to estimate the regression coefficients. This paper surveys the wide variety of fixed effects methods and their implementation in SAS, specifically, linear models with PROC GLM, logistic regression models with PROC LOGISTIC, models for count data with PROC GENMOD, and survival models with PROC PHREG.

INTRODUCTION

For many years, the most challenging task in statistics has been the effort to devise methods for making causal inferences from nonexperimental data. And within that project the most difficult problem is how to statistically control for variables that cannot be observed. For experimentalists, the solution to that problem is easy. Random assignment to treatment groups makes those groups approximately equal on *all* characteristics of the subjects, whether those characteristics are observable or unobservable. But in nonexperimental research, the classic way to control for potentially confounding variables is to measure them and put them in some kind of regression model. Without measurement, there is no control.

In this book, I describe a class of regression methods, called fixed effects models, that make it possible to control for variables that have not or cannot be measured. The basic idea is very simple: use each individual as his or her own control. For example, if you want to know whether marriage reduces recidivism among chronic offenders, compare an individual's arrest rate when he is married with his arrest rate when he is not married. Assuming that nothing else changes (a big assumption), the difference in arrest rates between the two periods is an estimate of the marriage effect for that individual. And if we average those differences across all persons in the population, we get an estimate of the average "treatment effect." This estimate controls for all stable characteristics of the offender. It controls both for easily measured variables like sex, race, ethnicity, and region of birth, as well as for more difficult variables like intelligence, parents' child-rearing practices, and genetic makeup. While it does not control for time-varying variables like employment status and income, these may be handled by the more conventional approach of measuring and putting them in a regression model.

There are two basic data requirements for using fixed effects methods. First, the dependent variable must be measured for each individual on at least two occasions. Those measurements must be directly comparable, that is, they must have the same meaning and metric. Second, the predictor variables of interest must change in value across those two occasions for some substantial portion of the sample. Fixed effects methods are pretty much useless for estimating the effects of variables that don't change over time, like race and sex. Of course, some statisticians argue that it makes no sense to talk about causal effects of such variables anyway (Sobel 2000).

SAS is an excellent computing environment for implementing fixed effects methods. Last year, SAS Publishing brought out my book *Fixed Effects Regression Methods for Longitudinal Data Using SAS*. Why is a whole book needed for fixed effects methods? First, rather different methods are needed for different kinds of dependent variables, whether quantitative, categorical, count, or event time. Second, for a specific kind of dependent variable, there are often two or more ways to implement the fixed effects approach, and we need to understand their differences and similarities. Third, and most challenging, special methods are needed (but not always available) when measured predictor variables are not "strictly exogenous," for example, when a dependent variable at one point in time may affect a predictor variable at a later point in time.

The term "fixed effects model" is usually contrasted with "random effects model". Unfortunately, this terminology is the cause of much confusion. In the classic view, a fixed effects model treats unobserved differences between individuals as a set of fixed parameters that can either be directly estimated, or partialed out of the estimating equations. In a random effects model, unobserved differences are treated as random variables with a specified probability distribution. If you consult the experimental design literature for explanations of the difference, you will find statements like the following:

> Common practice is to regard the treatment effects as fixed if those treatment levels used are the only ones about which inferences are sought . . . . If inferences are sought about a broader collection of treatment effects than those used in the experiment, or if the treatment levels are not selected purposefully . . . , it is common practice to regard the treatment effects as random (LaMotte 1983, pp.. 138-139).

Such characterizations are very unhelpful in a non-experimental setting, however, because they suggest that a random effects approach is nearly always preferable. Nothing could be further from the truth.

In a more modern framework (Wooldridge 2002), the unobserved differences are always regarded as random variables. Then, what distinguishes the two approaches is the structure of the correlations between the observed variables and the unobserved variables. In a random effects model, the unobserved variables are assumed to be uncorrelated with all the observed variables. In a fixed effects model, the unobserved variables are allowed to have any correlations whatever with the observed variables (which turns out to be equivalent to treating the unobserved variables as fixed parameters). Unless you allow for such correlations, you haven't really controlled for the effects of the unobserved variables. This is what makes the fixed effects approach so attractive.

Nevertheless, there are also some potentially serious disadvantages of a fixed effects approach. As already noted, a classic fixed effects approach will not produce any estimates of the effects of variables that don't change over time. Second, in some cases, fixed effects estimates may have substantially larger standard errors than random-effects estimates, leading to higher *p*-values and wider confidence intervals. The reason is simple. Random effects estimates use information both within and between individuals. Fixed effects estimates, on the other hand, use only within-individual differences, essentially discarding any information about differences between individuals. If predictor variables vary greatly across individuals but have little variation over time for each individual, then fixed effects estimates will be rather imprecise.

Why discard the between-individual variation? Because it is likely to be confounded with unobserved characteristics of the individuals. The idea is to get rid of the "contaminated" variation and use only the variation that produces approximately unbiased estimates of the parameters of interest. So, in statistical terms, we sacrifice efficiency in order to reduce bias. In non-experimental studies, I think this is often a good trade-off.

Another attractive thing about fixed effects methods is that software for implementing them is already widely available. For the basic linear models, for example, an ordinary least squares regression program like PROC GLM will usually suffice. More advanced linear models can be estimated with programs for doing structural equation modeling, like PROC CALIS. For logistic regression models, you can get by with a conventional logistic regression program for the two-period case. The multi-period case can be handled by doing conditional logistic regression, now available in PROC LOGISTIC. Fixed effects models for count data, can be estimated with conventional Poisson and negative binomial regression programs like PROC GENMOD. And finally, models for survival analysis can be estimated with a standard Cox regression program like PROC PHREG. In this paper I demonstrate how each of these procedures can be used to estimate fixed effects regression models

LINEAR MODELS

In this section, I consider fixed effects methods for data in which the dependent variable is measured on an interval scale and is linearly dependent on a set of predictor variables. We have a set of individuals (i=1,…, n), each of whom is measured at two or more points in time (t = 1,…, T).

Here's the notation. We let y_{it} be the dependent variable. We have a set of predictor variables that vary over time, represented by the vector $\mathbf{x}_{it}$, and another set of predictor variables $\mathbf{z}_i$ that do not vary over time. (If you're not comfortable with vectors, you can interpret these as single variables). Our basic model for y is

$$y_{it} = \mu_t + \beta\mathbf{x}_{it} + \gamma\mathbf{z}_i + \alpha_i + \varepsilon_{it} \qquad (1)$$

where μ_t is an intercept that may be different for each point in time, and β and γ are vectors of coefficients. The two "error" terms, α_i and ε_{it}, behave somewhat differently. There is a different ε_{it} for each individual at each point in time, but α_i only varies across individuals, not over time. We regard α_i as representing the combined effect on y of all unobserved variables that are constant over time. On the other hand, ε_{it} represents purely random variation at each point in time.

At this point, I make some rather strong assumptions about ε_{it}, namely, that each ε_{it} has a mean of zero, a constant variance (for all i and t), and is statistically independent of everything else (except for y). The assumption of zero

mean is not critical as it is only relevant for estimating the intercept. The constant variance assumption can sometimes be relaxed to allow for different variances for different *t*. Note, that the error term at any one point in time is independent of $\mathbf{x}_{it}$ at any *other* point in time, which means that $\mathbf{x}_{it}$ is *strictly exogenous*. This assumption may be relaxed in some situations, but the issues involved are neither trivial nor purely technical.

As for α_i, the traditional approach in fixed effects analysis is to assume that this term represents a set of *n* fixed parameters that can either be directly estimated or removed in some way from the estimating equations. As noted earlier, we'll take a more modern approach by assuming that α_i represents a set of random variables. Although we'll assume statistical independence of α_i and ε_{it}, we allow for *any* correlations between α_i and $\mathbf{x}_{it}$, the vector of time-varying predictors. And if we are not interested in γ, we can also allow for any correlations between α_i and $\mathbf{z}_i$. The inclusion of such correlations distinguishes the fixed effects approach from a random effects approach, and allows us to say that the fixed effects method "controls" for time-invariant unobservables. At this point, we don't need to impose any restrictions on the mean and variance of α_i.

THE TWO-PERIOD CASE

Estimation of the model in (1) is particularly easy when the variables are observed at only two points in time (*T*=2). The two equations are then

$$\begin{aligned} y_{i1} &= \mu_1 + \beta\mathbf{x}_{i1} + \gamma\mathbf{z}_i + \alpha_i + \varepsilon_{i1} \\ y_{i2} &= \mu_2 + \beta\mathbf{x}_{i2} + \gamma\mathbf{z}_i + \alpha_i + \varepsilon_{i2} \end{aligned} \qquad (2)$$

If we subtract the first equation from the second, we get the "first difference" equation:

$$y_{i2} - y_{i1} = (\mu_2 - \mu_1) + \beta(\mathbf{x}_{i2} - \mathbf{x}_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1}) \qquad (3)$$

which can be rewritten as

$$y_i^* = \mu^* + \beta\mathbf{x}_i^* + \varepsilon_i^* \qquad (4)$$

where the asterisks indicate difference scores. Notice that both α_i and $\gamma\mathbf{z}_i$ have been "differenced out" of the equation. Hence, we no longer have to be concerned about α_i and its possible correlation with $\mathbf{x}_i$*. On the other hand, we also lose the possibility of estimating γ. Since $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ are each independent of ε_{i1} and ε_{i2}, it follows that $\mathbf{x}_i$* is independent of ε_i*. That implies that one can get unbiased estimates of β by doing ordinary least squares (OLS) regression on the difference scores.

Let's try it on some real data. Our sample comes from the National Longitudinal Survey of Youth (NLSY) (Center for Human Resource Research 2002). Using a subset of a much larger sample, we have 581 children who were interviewed in 1990, 1992 and 1994. We will work with just three variables that were measured in each of the three interviews:

ANTI	antisocial behavior (scale ranges from 0 to 6)
SELF	self-esteem (scale ranges from 6 to 24)
POV	coded 1 if family is in poverty, otherwise 0.

At this point, we're going to ignore the observations in the middle year, 1992, and use only the data in 1990 and 1994. Our objective is to estimate a linear equation with ANTI as the dependent variable and SELF and POV as independent variables:

$$\text{ANTI}_t = \mu_t + \beta_1\text{SELF}_t + \beta_2\text{POV}_t + \alpha + \varepsilon_t \qquad t = 1, 2. \qquad (5)$$

By expressing the model in this way, we are assuming a particular direction of causality, specifically, that SELF and POV affect ANTI and not the reverse. We also assume that that the effects are contemporaneous (no lagged effects of SELF and POV). Both of these assumptions can be relaxed. Lastly, we assume that β_1 and β_2 are the same at both time points, an assumption that can also be relaxed.

I began by estimating equation (5) separately for each time point, using OLS regression with PROC REG

```
PROC REG DATA=my.nlsy;
 MODEL anti90=self90 pov90;
 MODEL anti94=self94 pov94;
RUN;
```

Results are shown in the first two columns Table 1. In both years, poverty is associated with higher levels of antisocial behavior while self esteem is associated with lower levels. The coefficients are quite similar across the two years.

Table 1. OLS Regression of Antisocial Behavior on Self-Esteem and Poverty

	1990	1994	Difference Score
Intercept	2.375 (.384)	2.888 (.447)	.209 (.063)
Self-Esteem	-.050 (.019)**	-.064 (.021)**	-.056 (.015)**
Poverty	.595 (.126)**	.547 (.148)**	-.036 (.128)
R^2	.05	.04	.02

Note: Standard errors in parentheses.
**p<.01

In neither of these two regressions are there any controls for time-invariant covariates, like sex and race. Rather than putting a bunch of variables in the equation, however, we can control for *all* time invariant covariates by doing the regression with difference scores. For each child and each variable, we subtract the 1990 value from the 1994 value, then regress the ANTI difference on the SELF difference and the POV difference. Since POV is a dummy variable, it might seem problematic to subtract one value from the other. But, in fact, dummy variables can be treated just like any other variables in this regard. Here's the code:

```
DATA diff;
  SET my.nlsy;
  antidiff=anti94-anti90;
  povdiff=pov94-pov90;
  selfdiff=self94-self90;
PROC REG DATA=diff;
  MODEL antidiff=selfdiff povdiff;
RUN;
```

Results are in the last column of Table 1. Although the equation was estimated in the form of difference scores, the coefficients can be interpreted as if we had estimated equation (5) directly. That is, they represent of the effects of each variable in a given year on the value of the dependent variable in the same year. For self-esteem, the estimated coefficient is in between the coefficients for the two separate years and still highly significant. For poverty, on the other hand, the coefficient is dramatically smaller and no longer statistically significant.

It's fairly common for fixed estimates to vary markedly from those produced by other methods. In this case, one possible interpretation is that the estimates for the poverty effect in the regressions for the separate years were spurious, reflecting the correlation between poverty and some time-invariant variables that affected antisocial behavior. However, whenever the application of a fixed effect method reduces a coefficient to nonsignificance, it's a good idea to check the standard errors. As already mentioned, standard errors for fixed effects coefficients may be substantially larger than those for other methods, especially when the predictor variable has little variation over time. In fact, most of the variation in poverty is between girls, with only about 24 percent of the girls moving into or out of poverty between 1990 and 1994. Nevertheless, the standard error of the poverty coefficient in the difference score model is about the same as the standard error in 1990 and noticeably smaller than the standard error in 1994. The conclusion, then, is that insufficient information is not a problem here. There seems to be a real and substantial decline in the magnitude of the poverty effect when time-invariant variables are controlled.

DUMMY VARIABLE METHOD FOR THREE OR MORE OBSERVATIONS PER INDIVIDUAL

When there are three or more observations per individual, a different method is needed. One such method, the dummy variable method, requires a data set with a rather different structure: one record for each observation for each individual. For the NLSY data, for example, the working data set must have three records for each of the 581 children, for a total of 1743 records. The time-varying variables have the same variable names on each record but different values. For any time-invariant variables, their values are simply replicated across the multiple records for each individual. There should also be an ID variable with a common value for all the records for each individual.

Lastly, there should be a variable distinguishing the different observations for each individual. For the NLSY data, for example, the variable TIME had values of 1, 2 and 3, corresponding to 1990, 1992 and 1994. Table 2 shows the first 15 records of this data set, corresponding to the first three persons.

Table 2. Data Set With Three Observations Per Person (First Three Persons).

ID	TIME	ANTI	SELF	POV	GENDER
1	1	1	21	1	1
1	2	1	24	1	1
1	3	1	23	1	1
2	1	0	20	0	1
2	2	0	24	0	1
2	3	0	24	0	1
3	1	5	21	0	0
3	2	5	24	0	0
3	3	5	24	0	0

To implement the method, it's necessary to construct a set of dummy variables to distinguish the individuals in the data set. In our example, that means 580 dummy variables to represent the 581 children. This is conveniently done with PROC GLM by specifying ID as a CLASS variable.

```
PROC GLM DATA=my.nlsy3;
 CLASS id time;
 MODEL anti=self pov time id /SOLUTION;
RUN;
```

Results are shown in the left-hand panel of Table 3. Only the coefficients for the first nine dummy variables are reported. The coefficients for the dummy variables created from the ID variable are actually estimates of the α_i in eq. (1), under the constraint that one of them is equal to 0.

Table 3. Regression of Antisocial Behavior on Self-Esteem and Poverty, Dummy Variable Method

	Fixed Effects			Conventional OLS		
	Coefficient	Std. Err.	*p*	Coefficient	Std. Err.	*p*
SELF	-.055	.010	0.00	-.067	.011	0.00
POV	.112	.093	0.23	.518	.079	0.00
TIME_2	.044	.059	0.45	.051	.090	0.58
TIME_3	.211	.059	0.00	.223	.091	0.01
ID_2	-.887	.819	0.28			
ID_3	4.131	.811	0.00			
ID_4	1.057	.819	0.20			
ID_5	-.536	.819	0.51			
ID_6	.040	.820	0.96			
ID_7	2.170	.821	0.01			
ID_8	.910	.820	0.27			
ID_9	-.276	.819	0.74			

For comparison, the right-hand panel of Table 3 gives the OLS estimates of the coefficients without the inclusion of the 580 dummy variables. As we saw in the two-period case, the big difference in the results for the two methods is that the coefficient for POV is much larger for conventional OLS, and highly significant. Thus, the apparent effect of poverty on self-esteem disappears when we adjust for all between-person differences and focus only on within-person changes. It's also of some interest to compare the standard errors. The standard error for the POV coefficient is larger for the fixed effects estimate, a fairly typical result arising from the non-use of the between-person variation. On the other hand, for the coefficients for SELF and the two TIME dummies, the fixed effects standard errors are actually smaller than those for conventional OLS. Why the difference? It's all a matter of the relative magnitudes of within- and between-person variation. For POV, 70 percent of the variation is between persons, while for SELF the

figure is only 53 percent.[1] For the TIME dummies, all of the variation is within person and none is between. The best situation for a fixed effects analysis is when all of the variation on a time-varying predictor is within persons.

The problem with the dummy variable method is that the computational requirement of estimating coefficients for all the dummy variables can be quite burdensome, especially in large samples where it may be beyond the capacity of the software or the machine memory. Fortunately, there is an alternative algorithm—the mean deviation method—that produces exactly the same results. It doesn't give estimates for the dummy variable coefficients, but those are rarely of interest anyway.

The mean deviation algorithm works like this. For each person and for each time-varying variable (both response and predictor variables), we compute the means over time for that person:

$$\bar{y}_i = \frac{1}{n_i}\sum_t y_{it}$$

$$\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_t \mathbf{x}_{it}$$

where n_i is the number of measurements for person i. Then we subtract the person-specific means from the observed values of each variable:

$$y^*_{it} = y_{it} - \bar{y}_i$$
$$\mathbf{x}^*_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$$

Finally, we regress y^* on $\mathbf{x}^*$, plus variables to represent the effect of time. This is sometimes called a "conditional" method because it conditions out the coefficients for the fixed effects dummy variables.

You can construct the deviation scores yourself and then use an ordinary regression program to estimate the coefficients. However, it's much easier to let SAS do it automatically with the ABSORB statement in PROC GLM.

```
PROC GLM DATA=my.nlsy3;
 ABSORB id;
 CLASS time;
 MODEL anti=self pov time /SOLUTION;
RUN;
```

This produces exactly the same coefficients, standard errors and *p*-values as those shown in the left panel of Table 3, although no estimates are given for the ID variable.

One characteristic of this method is the inability to estimate coefficients for time-invariant predictors. This is evident from the fact that subtracting the person-specific mean of a time-invariant predictor from the individual values (which are the same at all points in time) yields a value of 0 for all persons. Keep in mind, however, that we are still controlling all time invariant predictors even though they drop out of the equation. It's also possible, and often useful, to include interactions between time-invariant predictors and time-varying predictors (including time itself).

LOGISTIC MODELS FOR DICHOTOMOUS OUTCOMES

MODELS FOR DATA WITH TWO OBSERVATIONS PER PERSON.

I now consider the situation in which the response variable is a dichotomy and there are exactly two observations for each individual. As before, we let y_{it} be the value of the response variable for individual i on occasion t, but now y is constrained to have values of either 0 or 1. In this section, t=1 or 2. Let p_{it} be the probability that y_{it}=1. It is convenient to assume that the dependence of p_{it} on possible predictor variables is described by a logistic regression model

[1] These numbers can be obtained by running an analysis of variance with each variable as the dependent variable and the ID variable as a categorical predictor.

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \mu_t + \beta x_{it} + \gamma z_i + \alpha_i \qquad (6)$$

where z_i is a column vector of variables that describe the individuals but do not vary over time, and x_{it} is a column vector of variables that vary both over individuals and over time for each individual. In this equation, μ_t is an intercept that is allowed to vary with time, and β and γ are row vectors of coefficients. As before, α_i represents all differences between persons that are stable over time and not otherwise accounted for by z_i. Again, we regard these as fixed parameters, one per person. Additionally, we assume that for a given individual *i* (and, hence, a given value of α_i), y_{i1} and y_{i2} are independent.

Applying some algebra to equation 6 and making use of the independence assumption, it can be shown that

$$\log\left(\frac{\Pr(y_{i1}=0, y_{i2}=1)}{\Pr(y_{i1}=1, y_{i2}=0)}\right) = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}) \qquad (7)$$

Thus, as we found for the linear model, both z_i and α_i have been "differenced" out of the equation. This result suggests the following method for estimating the parameters:

- Eliminate all individuals who do not change on the response variable.
- Create difference scores for all the time-varying predictors.
- Use maximum likelihood to estimate the logistic regression predicting y_{i2}, with the difference scores as predictor variables.

This procedure is a form of *conditional logistic regression*.

Here is an example. The sample consists of 1151 girls from the National Longitudinal Survey of Youth who were interviewed annually for nine years, beginning in 1979. For this initial example, we'll only use data from year 1 and year 5. The response variable POV has a value of 1 if the girl's household was in poverty (as defined by U.S. federal standards) in each of the years, otherwise 0. The predictor variables are:

AGE	Age in years at the first interview
BLACK	1 if respondent is black, otherwise 0
MOTHER	1 if respondent currently had a least one child, otherwise 0
SPOUSE	1 if respondent is currently living with a spouse, otherwise 0
INSCHOOL	1 if respondent is currently enrolled in school, otherwise 0
HOURS	Hours worked during the week of the survey

The first two variables are time-invariant while the last four may differ at each interview.
The data set MY.TEENPOV has one record for each of the 1151 respondents, with different variable names for the same variable measured in different years.

To do the logistic regression analysis, the first step is to create a new data set that excludes those girls whose poverty status was the same in years 1 and 5, and includes new variables that are differences between year 5 values and year 1 values.

```
DATA teendif;
  SET my.teenpov;
  IF pov1=pov5 THEN DELETE;
  mother=mother5-mother1;
  spouse=spouse5-spouse1;
  inschool=inschool5-inschool1;
  hours=hours5-hours1;
RUN;
```

Next, we estimate a logistic regression with POV5 as the dependent variable, and difference scores and time-invariant predictors as independent variables:

```
PROC LOGISTIC DATA=teendif DESC;
  MODEL pov5=mother spouse inschool hours black age;
RUN;
```

Table 4 gives the results. Although the time-varying predictors are expressed as difference scores, their coefficients should be interpreted as they appear in equation (6), that is, as the effect of the value of the variable in a given year on the probability of poverty in that same year. Thus, the odds-ratio for MOTHER tells us that girls with children had twice the odds of being in poverty as girls without children (net of other variables). On the other hand, for girls living with husbands, the odds of poverty was only 35 percent as large as the odds for those not living with husbands. Each additional hour of work per week reduced the odds of poverty by 100(1-.967) = 3.3 percent.

Table 4. LOGISTIC Output for Regression on Difference Scores

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	4.8993	1.6438	8.8829	0.0029
mother	1	0.7436	0.2538	8.5862	0.0034
spouse	1	-1.0317	0.2918	12.5014	0.0004
inschool	1	0.3394	0.2178	2.4287	0.1191
hours	1	-0.0339	0.00623	29.7027	<.0001
black	1	-0.5263	0.2164	5.9154	0.0150
age	1	-0.2577	0.1029	6.2739	0.0123

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
mother	2.103	1.279	3.459
spouse	0.356	0.201	0.631
inschool	1.404	0.916	2.152
hours	0.967	0.955	0.978
black	0.591	0.387	0.903
AGE	0.773	0.632	0.945

The coefficients (and odds ratios) for BLACK and AGE must be interpreted somewhat differently. According to equation (7), as time-invariant predictors these variables shouldn't even be in the model. In fact, they represent interactions between time-invariant predictor variables and time itself, so that the rate of change in the odds of poverty depends on the value of these variables. More specifically, for a girl whose predictor variables did not change from year 1 to year 5, the change in the log-odds of poverty over the five-year period can be expressed as

4.8993 - .5263*BLACK - .2577*AGE.

Thus, for a 14-year-old girl who was not black and who did not change on any of the other predictors, the predicted change in the log-odds is + 1.29. Equivalently, her odds of being in poverty increase by a factor of exp(1.29)=3.63. We conclude that blacks and girls who were older at year 1 had a lower rate of increase in poverty.

LOGISTIC MODELS FOR THREE OR MORE OBSERVATIONS PER PERSON

When individuals in the sample have three or more observations, we can't use the simple method of doing a logistic regression on the persons who change (with difference scores as predictors). In the case of linear models, we solved this problem by expressing all variables as deviations from the person-specific means. In the case of dichotomous outcomes, there is an analogous method that can be implemented with PROC LOGISTIC (release 9.0 and later)[2].

Earlier we saw that one way to estimate a fixed effects linear model in the multiple observation case was to structure the data with one observation per individual per occasion and then compute an OLS regression with dummy variables

[2] Conditional logistic regression requires the STRATA statement, which was first implemented in release 9.0 of SAS. For earlier releases, conditional logistic regression can be accomplished with PROC PHREG using the methods described in Allison (1999).

for all individuals (except one). However, the device of using dummy variables does not work for logistic regression. The coefficients are generally biased upward and the test statistics will also be incorrect. How come?

This is an example of a general problem called the *incidental parameters problem* (Kalbfleisch and Sprott 1970) that arises in certain applications of maximum likelihood estimation. The justification for maximum likelihood estimators is usually asymptotic, which means that it's based on how the estimators behave as the sample gets large. However, the validity of that justification depends on the presumption that the number of parameters remains constant as the sample gets larger. For longitudinal data, that works just fine if the number of individuals remains constant but the number of observations per individual gets larger. But if the number of individuals is getting larger while the number of time points remains constant, then the number of parameters in a fixed effects model (including coefficients of the dummy variables) is increasing at the same rate as the sample size. This is not a problem with linear models and (somewhat surprisingly) for the Poisson models. But it is a serious problem with logistic regression and many other nonlinear regression models. The biases are greatest when, as in the previous section, the number of time points per individual is small.

The solution to this problem is to do *conditional maximum likelihood* (Chamberlain 1980), which we already employed in the two-period case. How can we implement conditional logistic regression in PROC LOGISTIC? In the previous section, we estimated a conditional logistic regression model for poverty in years 1 and 5 of a five-year series. Now let's look at all five years together. The first thing we must do is restructure the data so that there is one record per person-year instead of one record per person:

```
DATA teenyrs5;
  SET my.teenpov;
  ARRAY pv(*) pov1-pov5;
  ARRAY mot(*) mother1-mother5;
  ARRAY spo(*) spouse1-spouse5;
  ARRAY ins(*) inschool1-inschool5;
  ARRAY hou(*) hours1-hours5;
  DO year=1 TO 5;
    pov=pv(year);
    mother=mot(year);
    spouse=spo(year);
    inschool=ins(year);
    hours=hou(year);
    OUTPUT;
  END;
  KEEP id year black age pov mother spouse inschool hours;
RUN;
```

This data step produces 5755 observations, five for each of the 1151 girls. Now we're ready to run PROC LOGISTIC to estimate the first model:

```
PROC LOGISTIC DATA=teenyrs5 DESC;
  CLASS year /PARAM=REF;
  MODEL pov = year mother spouse inschool hours;
  STRATA id;
RUN;
```

The CLASS statement declares YEAR to be a categorical variable, with the highest year (year 5) being the reference category. The STRATA statement says that each girl is a separate stratum, which has the consequence of grouping together the five observations for each girl in the process of constructing the likelihood function.

Results in Table 5 are rather similar to those in Table 4, which was based on only two observations per person. In the "Analysis of Maximum of Likelihood Estimates" panel, we see that motherhood and school enrollment increase the risk of poverty while living with a husband and working more hours reduce the risk. The last panel gives the odds ratios. We see that motherhood increases the odds of poverty by an estimated 79 percent. Living with a husband cuts the odds approximately in half. Each additional hour of employment per week reduces the odds by about 2 percent. Keep in mind that these estimates control for *all* stable characteristics of the girls, including such things as race, intelligence, place of birth and parent's education.

Table 5. Conditional Logistic Regression Estimates Produced by LOGISTIC

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
year	1	1	-0.4025	0.1275	9.9615	0.0016
year	2	1	-0.0707	0.1185	0.3562	0.5506
year	3	1	-0.0675	0.1096	0.3793	0.5380
year	4	1	0.0303	0.1047	0.0836	0.7725
mother		1	0.5824	0.1596	13.3204	0.0003
spouse		1	-0.7478	0.1753	18.1856	<.0001
inschool		1	0.2719	0.1127	5.8157	0.0159
hours		1	-0.0196	0.00315	38.8887	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
year	1 vs 5	0.669	0.521	0.859
year	2 vs 5	0.932	0.739	1.175
year	3 vs 5	0.935	0.754	1.159
year	4 vs 5	1.031	0.840	1.265
mother		1.790	1.310	2.448
spouse		0.473	0.336	0.668
inschool		1.312	1.052	1.637
hours		0.981	0.975	0.987

Although models like this cannot include the main effects of time-invariant variables, they do allow for interactions between time-invariant variables and time-varying variables, including time itself. The next model, for example, includes the interaction between MOTHER and BLACK.

```
PROC LOGISTIC DATA=teenyrs5 DESC;
  CLASS year /PARAM=REF;
  MODEL pov = year mother spouse inschool hours mother*black;
  STRATA id;
RUN;
```

In Table 6, we see that the interaction is statistically significant at the .05 level. For nonblack girls, the effect of motherhood is to increase the odds[3] of poverty by a factor of exp(.9821)=2.67. For black girls, on the other hand, the effect of motherhood is to increase the odds of poverty by a factor of exp(.9821-.5989)= 1.47. Thus, motherhood has a larger effect on poverty status among nonblack girls than among black girls.

Table 6. Conditional Logistic Regression with Interaction

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
year	1	1	-0.3996	0.1276	9.8046	0.0017
year	2	1	-0.0677	0.1186	0.3260	0.5680
year	3	1	-0.0654	0.1097	0.3552	0.5512

[3] By default, PROC LOGISTIC does not report odds ratios for variables involved in an interaction. However, these can be requested with the EXPB option on the MODEL statement.

year	4	1	0.0304	0.1047	0.0843	0.7716
mother		1	0.9821	0.2529	15.0787	0.0001
spouse		1	-0.7830	0.1777	19.4224	<.0001
inschool		1	0.2671	0.1128	5.6084	0.0179
hours		1	-0.0192	0.00316	36.9396	<.0001
mother*black		1	-0.5989	0.2897	4.2748	0.0387

FIXED EFFECTS REGRESSION METHODS FOR COUNT DATA

Many response variables are counts of something: number of articles published by scientists, number of sex partners in the last year, number of arrests in a one-year period, number of students enrolling for a class, and so on. Some data analysts still treat count variables as continuous measures and apply ordinary linear regression. But that ignores two facts: the data are really discrete, and the distributions of count variables are typically highly skewed. For these reasons, it may be inappropriate to use models that assume normally distributed errors.

Nowadays, it's becoming increasingly popular to estimate Poisson regression models or negative binomial regression models, both of which are explicitly designed to model count data. In this section we'll see how to extend these count data methods to handle multiple observations per individual, with the inclusion of fixed effects to control for all stable predictor variables. Along the way, we'll revisit many of the issues that arose for dichotomous outcomes in Chapter 3, but the problems encountered there turn out to be less serious for count data models.

Let's begin by describing the example. The data consist of 346 manufacturing firms with yearly counts of patents received in each of the years 1975 to 1979. There is one record per firm, with variables PAT75-PAT79 containing the patent counts in the five years. As predictors we have the logarithm of research and development expenditures from 1970 to 1979 (LOGR70-LOGR79). There are also two time-invariant predictors: SIZE, which is the book value of the firm in 1972, and SCIENCE, an indicator variable equal to 1 if the firm is in the science sector, otherwise 0.

POISSON MODELS FOR COUNT DATA WITH TWO OBSERVATIONS PER INDIVIDUAL

When there are only two observations per individual, we previously saw that a linear or logistic fixed effects analysis could be done using simplified methods with conventional software. This is also true for count data. In fact, a fixed effects Poisson regression model can be estimated with an ordinary logistic regression program.

For the patent data, let's ignore the intervening years and focus only on 1975 and 1979. Let y_{i1} be the patent count for firm i in 1975 and y_{i2} the patent count in 1979. We assume that each of these variables has a Poisson distribution with parameter λ_{it}. That is, the probability that $y_{it} = r$ is given by

$$\Pr(y_{it} = r) = \frac{\lambda_{it}^{r} e^{-\lambda_{it}}}{r!}, \qquad r = 0, 1, 2, \ldots \tag{8}$$

Note that we are not assuming that there is a single Poisson distribution for the entire sample. Rather each firm's patent count is drawn from a different Poisson distribution whose parameter λ_{it} varies across both firms and time.

An important property of the Poisson distribution is that its mean and variance are equal, and both are equal to the Poisson parameter:

$$E(y_{it}) = \mathrm{var}(y_{it}) = \lambda_{it} \tag{9}$$

Next, we let λ_{it} be a loglinear function of the predictor variables

$$\log \lambda_{it} = \mu_t + \beta x_{it} + \gamma z_i + \alpha_i \tag{10}$$

As in earlier chapters, x_{it} represents the time-varying predictor variables, z_i denotes the time-invariant predictors, and α_i denotes the unobserved "fixed effects". The vector x_{it} includes the R & D expenditures in the current year t and in each of the preceding five years.

Our goal is to estimate the parameters in (10). To do this, we shall use conditional maximum likelihood, the same method used previously to estimate the fixed effects logistic regression model. Consider the distribution of y_{i2} conditional on the total event count for the two time periods combined, denoted by $w_i = y_{i1} + y_{i2}$. It can be shown that

$y_{i2}|w_i \sim B(p_i, w_i)$. That is, conditional on the total count, y_{i2} has a binomial distribution with parameters p_i and w_i, where

$$p_i = \frac{\lambda_{i2}}{\lambda_{i2} + \lambda_{i1}}. \quad (11)$$

It follows, after a bit of algebra, that

$$\log\left(\frac{p_i}{1-p_i}\right) = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}). \quad (12)$$

Thus, we have converted our Poisson regression model into a logistic regression model in which the predictor variables are difference scores for the original predictors. Note that, as in earlier applications, both α_i and γz_i drop out of equation (12).

To implement this conditional approach in SAS, we may use any SAS procedure that does logistic regression for grouped data. Here's how to do it in GENMOD. First, we create a new data set that contains the total count for each firm and the difference scores for the research and development variables:

```
DATA patents;
  SET my.patents;
  total=pat75+pat79;
  rd_0=logr79-logr75;
  rd_1=logr78-logr74;
  rd_2=logr77-logr73;
  rd_3=logr76-logr72;
  rd_4=logr75-logr71;
  rd_5=logr74-logr70;
RUN;
```

RD_0 is the difference score for the same years in which the patents were counted, while RD_1 through RD_5 are difference scores for lags of one to five years.

Let's first estimate a model with the lagged R&D measures as covariates:

```
PROC GENMOD DATA=patents;
  MODEL pat79/total = rd_0-rd_5 / DIST=B;
RUN;
```

Note that the dependent variable is expressed with the *events/trials* syntax, which tells SAS that PAT79 events occurred out of a possible TOTAL. DIST=B specifies that PAT79 has a binomial distribution whose default link function is logit (i.e., logistic). Results are shown in Table 7.

Table 7. Conditional Poisson Regression Model for Patents

Model Information

Data Set	WORK.PATENTS
Distribution	Binomial
Link Function	Logit
Response Variable (Events)	pat79
Response Variable (Trials)	total
Observations Used	300
Number Of Events	11107
Number Of Trials	23865
Invalid Response Values	46

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	293	949.3031	3.2399
Scaled Deviance	293	949.3031	3.2399
Pearson Chi-Square	293	890.2903	3.0385
Scaled Pearson X2	293	890.2903	3.0385
Log Likelihood		-16458.7718	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.2225	0.0178	-0.2573	-0.1876	156.50	<.0001
rd_0	1	0.5214	0.0844	0.3561	0.6868	38.19	<.0001
rd_1	1	-0.2067	0.1129	-0.4280	0.0146	3.35	0.0671
rd_2	1	-0.1179	0.1110	-0.3355	0.0996	1.13	0.2880
rd_3	1	0.0601	0.0958	-0.1277	0.2478	0.39	0.5305
rd_4	1	0.1806	0.0900	0.0042	0.3569	4.03	0.0448
rd_5	1	-0.0932	0.0690	-0.2284	0.0420	1.83	0.1765
Scale	0	1.0000	0.0000	1.0000	1.0000		

Under “Model Information” we see that 46 firms had “invalid response values”. These are firms that had 0 patents in both 1975 and 1979, so their total for the two years was also 0. Of course, the binomial distribution is undefined when the number of trials is 0, which is why these firms are excluded. This points out a more general characteristic of Poisson regression that extends to the next section as well. Whenever you condition on the total count, those cases that have a total count of zero are effectively removed from the likelihood function. If the total is 0, then each component must also be 0, leaving no within-individual variability to analyze.

In the next panel, “Criteria for Goodness of Fit”, we see that both the deviance and the Pearson chi-square statistics are more than three times their degrees of freedom. For a good fitting model, these statistics should be close to their degrees of freedom. The magnitude of these ratios suggests that there is a problem with *overdispersion* that really needs to be dealt with.

Examining the parameter estimates and their associated statistics, we see that RD_0, the contemporaneous measure of research and development expenditures, has a highly significant effect on the patent count, with a coefficient of .5214. To interpret this, keep in mind that both the dependent variable (expected number of patents) and the independent variable (research and development expenditures) are logged. In that case, we can say that a one-*percent* increase in R & D expenditures is associated with a .52 percent increase in the expected number of patents in the same year, controlling for the lagged R & D measures. The effects of the lagged measures are much smaller.

Now let’s deal with the overdispersion problem. The big danger with overdispersion is that the standard errors may be underestimated, leading to chi-squares that are too large and *p*-values that are too low. There are several possible solutions to this problem, one of which is to formulate and estimate a model that directly builds in the overdispersion. But a simpler, less elegant, approach is to correct the standard errors and chi-squares based on the goodness-of-fit ratios that alerted us to the problem. In PROC GENMOD, this is accomplished by using the DSCALE or PSCALE options on the MODEL statement. For example,

```
PROC GENMOD DATA=patents;
  MODEL pat79/total = rd_0-rd_5 / DIST=B DSCALE;
RUN;
```

The DSCALE option uses the deviance chi-square to make the adjustment while PSCALE uses the Pearson chi-square. The adjustment is very simple: Calculate the square root of the ratio of the chi-square statistic to its degrees of freedom. In Table 8, this number is reported as the “Scale” parameter in the next-to-last line. All standard errors

are then multiplied by the scale parameter, which in turn attenuates the chi-squares and the *p*-values, as shown in Table 8.

Table 8. Conditional Poisson Regression Model With Overdispersion Adjustment

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.2225	0.0320	-0.2852	-0.1597	48.30	<.0001
rd_0	1	0.5214	0.1519	0.2238	0.8191	11.79	0.0006
rd_1	1	-0.2067	0.2032	-0.6050	0.1916	1.03	0.3091
rd_2	1	-0.1179	0.1998	-0.5095	0.2736	0.35	0.5550
rd_3	1	0.0601	0.1724	-0.2779	0.3980	0.12	0.7275
rd_4	1	0.1806	0.1620	-0.1369	0.4980	1.24	0.2649
rd_5	1	-0.0932	0.1241	-0.3365	0.1501	0.56	0.4527
Scale	0	1.8000	0.0000	1.8000	1.8000		

NOTE: The scale parameter was estimated by the square root of DEVIANCE/DOF.

When this is done for the patent data, we find that only RD_0 retains its statistical significance, and even for this variable the chi-square is greatly reduced. Note also that the coefficients are not modified at all by this overdispersion correction. Other approaches to overdispersion—like estimating a negative binomial model—may produce different coefficient estimates.

POISSON MODELS FOR DATA WITH MORE THAN TWO OBSERVATIONS PER INDIVIDUAL

When there are more than two observations per individual, estimation of a fixed effects Poisson model in SAS is not so straightforward. Let's extend the example of the last section by analyzing annual patent counts from 1975 to 1979, with each count denoted by y_{it}. As before, we assume that y_{it} has a Poisson distribution given by (8) with parameter λ_{it}, and we let λ_{it} be the loglinear function of the predictor variables given in (10).

There are two approaches to estimation, conditional maximum likelihood and unconditional maximum likelihood. In conditional ML, the likelihood function is conditioned on the total count for each individual, thereby eliminating the fixed effects (α_i). The resulting conditional likelihood (Cameron and Trivedi 1998) is proportional to

$$\prod_i \prod_t \left(\frac{\exp(\mu_t + \beta \mathbf{x}_{it})}{\sum_s \exp(\mu_s + \beta \mathbf{x}_{is})} \right)^{y_{it}} \qquad (13)$$

SAS has no procedure that is explicitly designed to maximize this likelihood. Happily, it turns out that, unlike logistic regression, unconditional maximum likelihood for the Poisson model always produces identical results to conditional maximum likelihood (Cameron and Trivedi 1998). The latter is accomplished by estimating a conventional Poisson regression model with dummy variables for all the firms (less one). The first step is to restructure the data so that there is one record for each firm year:

```
DATA patents2;
  SET my.patents;
  ARRAY pat (*) pat75-pat79;
  ARRAY logr (*) logr70-logr79;
  id=_N_;
  sumpat=pat79+pat78+pat77+pat76+pat75;
  IF sumpat NE 0 THEN DO t=1 TO 5;
     j=t+5;
     patent=pat(t);
    rd_0=logr(j);
```

```
    rd_1=logr(j-1);
    rd_2=logr(j-2);
    rd_3=logr(j-3);
    rd_4=logr(j-4);
    rd_5=logr(j-5);
    OUTPUT;
  END;
RUN;
```

Each record contains a patent count, a contemporaneous R & D value, and five lagged values of R & D, along with any other variables already in the MY.PATENTS data set. The new data set has 1620 observations. Once the new data set has been constructed, estimation with PROC GENMOD is straightforward:

```
PROC GENMOD DATA=patents2;
  CLASS t id;
  MODEL patent = rd_0-rd_5 t id / DIST=POISSON PSCALE;
  RUN;
```

Results are shown in Table 9.

Table 9. Unconditional Poisson Estimates for Five Years of Patent Counts

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr>ChiSq
Intercept	1	2.8057	0.2688	2.2789	3.3324	108.99	<.0001
rd_0	1	0.3222	0.0667	0.1915	0.4529	23.35	<.0001
rd_1	1	-0.0871	0.0707	-0.2257	0.0514	1.52	0.2176
rd_2	1	0.0786	0.0650	-0.0488	0.2060	1.46	0.2267
rd_3	1	0.0011	0.0601	-0.1168	0.1189	0.00	0.9859
rd_4	1	-0.0046	0.0549	-0.1123	0.1030	0.01	0.9327
rd_5	1	0.0026	0.0468	-0.0892	0.0944	0.00	0.9556
t 6	1	0.1980	0.0222	0.1545	0.2415	79.56	<.0001
t 7	1	0.1554	0.0220	0.1124	0.1985	50.11	<.0001
t 8	1	0.1580	0.0210	0.1168	0.1992	56.39	<.0001
t 9	1	0.0409	0.0203	0.0012	0.0806	4.08	0.0434
t 10	0	0.0000	0.0000	0.0000	0.0000	.	.

In this model we have corrected for overdispersion by using the PSCALE option on the MODEL statement (which adjusts the standard errors based on the Pearson chi-square). Again it is quite clear that lagged measures of R & D expenditures contribute little, if anything, to the prediction of patent counts beyond the contemporaneous measure.

Another, and possibly better, way to deal with overdispersion is to estimate a negative binomial rather than a Poisson model. As in the case of the Poisson, this is most easily accomplished by estimating an unconditional model, say, with PROC GENMOD. See my book for details.

FIXED EFFECTS METHODS FOR SURVIVAL ANALYSIS

In both the social and biomedical sciences, there is a great deal of interest in regression models for predicting the occurrence and timing of events. Statistical methods for modeling events are often called *survival analysis* because they were originally developed by biostatisticians to analyze the occurrence of deaths. But I prefer the term *event history analysis* because it more aptly expresses the generality of these methods, and because it is particularly appropriate for modeling *repeated events*, which are the main focus of this section.

As usual, we begin with an example. In the 1995 National Survey of Family Growth (NSFG), a representative sample of American women was asked to report information on the births of all children ever born to them. In the version of the data used here, 6,911 women reported on 14,932 live births. For each of these births, I calculated a birth interval, labeled DUR: the length of time (in months) from the current birth to the next birth, or until the interview date if no subsequent birth was observed. As potential predictors of these birth intervals, there is information on several

variables that characterize the current birth:

PREGORDR	Order of the birth (1 through 15)
MARRIED	1 if married at the time of the birth, otherwise 0
AGE	Mother's age (in years) at birth
PASST	1 if delivery was paid for, in part, by public assistance funds, otherwise 0
NOBREAST	1 if mother did not breast feed baby, otherwise 0
LBW	1 if low birth weight, otherwise 0
CAESAR	1 if birth was by Caesarian section, otherwise 0
MULTIPLE	1 if more than one baby born, otherwise 0

There is also a variable COLLEGE equal to 1 if the woman had some college education (at the time of the interview), otherwise 0, and a variable BIRTH equal to 1 if the interval ended in another birth or 0 if the interval was terminated by the interview (a right censored interval). Finally, there is a variable CASEID which is an ID number that is common to all the birth intervals for the same woman. The goal is to estimate a regression model predicting the length of birth interval.

COX REGRESSION

The most popular method for survival analysis is Cox regression, which consists of the *proportional hazards model* and the *partial likelihood* method for estimating that model. Before we discuss fixed effects analysis, it's essential to review this method.

Rather than directly modeling the length of the interval, the dependent variable in Cox regression is the *hazard* or instantaneous likelihood of event occurrence. For repeated events, the hazard may be defined as follows. Let $N_i(t)$ be the number of events that have occurred to individual i by time t. The hazard for individual i at time t is given by

$$h_i(t) = \lim_{\Delta t \to 0} \frac{\Pr[N_i(t+\Delta t) - N_i(t) = 1]}{\Delta t} \qquad (14)$$

In words, this equation says to consider the probability of one additional event in some small interval of time Δt. Then form the ratio of this probability to Δt, and take the limit of this ratio as Δt goes to 0. For repeated events, the hazard function is also known as the intensity function.

The next step is to model the hazard as a function of the predictor variables. Letting $h_{ik}(t)$ be the hazard for the k'th event for individual i, a proportional hazards model is given by

$$\log h_{ik}(t) = \mu(t - t_{i(k-1)}) + \beta x_{ik} \qquad (15)$$

where x_{ik} is a column vector of predictor variables that may vary across individuals and across events, β is a row vector of coefficients, $t_{i(k\text{-}1)}$ is the time of the (k-1)'th event, and $\mu(.)$ is an unspecified function. In this model, the hazard of an event depends on the time since the most recent event.

The method of partial likelihood makes it possible to estimate β without specifying anything about the function μ. For details on how this is accomplished, see Allison (1995). In SAS, partial likelihood is implemented with PROC PHREG. Here's a program for estimating the model in (15), *without* incorporating fixed effects:

```
PROC PHREG DATA=my.nsfg;
  MODEL dur*birth(0)=pregordr age married passt
    nobreast lbw caesar multiple college / TIES=EFRON;
RUN;
```

In the MODEL statement, the left-hand side of the equation is expressed as DUR*BIRTH(0), which is necessary to allow for the fact that many of the intervals are terminated by the interview rather than by another birth. In event history terminology, these are called *censored* intervals. The variable BIRTH indicates whether or not an interval is censored, and the number in parentheses (in this case 0) gives the value of the variable that corresponds to censored cases. The TIES=EFRON option requests a slight technical change in the estimation method that I recommend for routine use. See Allison (1995) for details.

In Table 10, we see that 6,911 of the birth intervals were censored. That's not surprising because the data collection method implies that each woman's last interval was terminated by the interview. Looking at the "Analysis of Maximum Likelihood Estimates", we find that all the variables but one (low birth weight) have highly significant effects on the hazard for a subsequent birth. Increased hazards are associated with being married or on public assistance. All the other variables have negative signs.

Table 10. Cox Regression Estimates for a Conventional Model

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
14932	8021	6911	46.28

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
pregordr	1	-0.16434	0.01150	204.0833	<.0001	0.848
age	1	-0.06565	0.00306	461.2265	<.0001	0.936
married	1	0.22320	0.02867	60.6010	<.0001	1.250
passt	1	0.13824	0.02868	23.2324	<.0001	1.148
nobreast	1	-0.27190	0.02332	135.9444	<.0001	0.762
lbw	1	-0.00246	0.04204	0.0034	0.9533	0.998
caesar	1	-0.11706	0.03054	14.6912	0.0001	0.890
multiple	1	-0.70661	0.14257	24.5635	<.0001	0.493
college	1	-0.20844	0.02598	64.3778	<.0001	0.812

Unfortunately, there's a potential problem with these results. Sixty-nine percent of the women contributed at least two birth intervals to the data set, and it's reasonable to suspect that there would be some dependence among these repeated observations. In particular, it's natural to suppose that some women have persistently short birth intervals while others have persistently long intervals. The failure to address this dependence could lead to serious underestimates of the standard errors and *p*-values. Beginning with release 8.1, PHREG includes an option called COVSANDWICH that makes it easy to correct for dependence when there are repeated observations. An alternative correction is to estimate a fixed effects model, which we now consider.

COX REGRESSION WITH FIXED EFFECTS

Now we're ready to introduce fixed effects into the Cox regression model. As usual, this makes it possible to control for all stable predictor variables, while at the same time addressing the problem of dependence among the repeated observations. As in earlier fixed effects models, α_i represents the combined effects of all stable covariates:

$$\log h_{ik}(t) = \mu(t - t_{i(k-1)}) + \beta x_{ik} + \alpha_i \qquad (16)$$

How can we estimate equation (16) for our birth interval data? One obvious possibility is to put dummy variables in the model for all women (except one). This method worked well for the Poisson and negative binomial regression models discussed earlier, but it runs into serious difficulties here. First, there is the practical problem of putting 6,910 dummy variables into a PHREG model. In principle, such computational difficulties could be solved by using Greene's (2001) algorithms, but these are not currently available in any commercial software.

The more fundamental difficulty is the potential bias introduced by estimating so many "incidental parameters." In previous chapters, we saw that this bias could be quite serious for logistic regression models, but not for Poisson or negative binomial models. Elsewhere (Allison 2002), I've shown that Cox regression is more like logistic regression in this regard. When the average number of intervals per person is less than three, regression coefficients are inflated by approximately 30 to 90 percent, depending on the level of censoring (a higher proportion of censored cases

produces greater inflation).

Fortunately, there is a simple alternative method that does the job very well. It's similar to the conditional likelihood methods used for logistic regression in that the coefficients for the dummy variables are not actually estimated but are eliminated from the likelihood function. First we modify equation (16) by defining

$$\mu_i(t - t_{i(k-1)}) = \mu(t - t_{i(k-1)}) + \alpha_i$$

which yields

$$\log h_{ik}(t) = \mu_i(t - t_{i(k-1)}) + \beta x_{ik} \qquad (17)$$

In this equation, the fixed effect α_i has been absorbed into the unspecified function of time, which is now allowed to vary from one individual to another. Thus, each individual is allowed to have her own hazard function, which is considerably less restrictive than allowing each individual to have her own constant.

Model (17) can be estimated by partial likelihood using the well-known method of stratification. Stratification allows different subgroups to have different baseline hazard functions, while constraining the coefficients to be the same across subgroups. It is accomplished by constructing a partial likelihood function for each subgroup, multiplying those likelihood functions together, and then maximizing the resulting likelihood function with respect to the coefficient vector β. In PHREG, stratification is implemented with the STRATA statement. Here's how it's done for the birth interval data:

```
PROC PHREG DATA=my.nsfg NOSUMMARY;
  MODEL dur*birth(0)= pregordr age married passt nobreast lbw
     caesar multiple college /TIES=EFRON;
  STRATA caseid;
RUN;
```

The statement STRATA CASEID creates a separate stratum for each value of CASEID, which means a separate stratum for each of the 6,911 women. That may seem like an enormous number of strata, but PHREG handles it with ease. The NOSUMMARY option is really optional, but it's strongly advised. If you don't include it, the output contains a line for each stratum, reporting the numbers of cases and events for that stratum.

The results in Table 11 show some noteworthy differences from those in Table 10. First, there's nothing reported for COLLEGE. Like most of our fixed effects methods, we can't estimate coefficients for variables that do not vary within person. Moving upward from COLLEGE, we see that the effect of a multiple birth is about the same as the previous estimates. But the coefficient for CAESAR is somewhat attenuated and no longer statistically significant. Low birth weight was previously far from statistically significant, but here the *p*-value is less than .01. The hazard ratio for LBW tells us that a low birth weight is associated with a 21 percent reduction in the hazard for a subsequent birth. The effect of breast feeding is attenuated, both in magnitude and significance. Public assistance was previously highly significant, but here it's not significant at all. The effect of marital status is about the same. Age is no longer statistically significant. On the other hand, the effect of pregnancy order is *much* greater, both in magnitude and statistical significance. Each additional birth is associated with about a 50 percent reduction in the hazard for a subsequent birth.

Table 11. Cox Regression with Fixed Effects Via Stratification

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
pregordr	1	-0.71663	0.03372	451.7316	<.0001	0.488
age	1	0.00818	0.01125	0.5285	0.4672	1.008
married	1	0.18307	0.06958	6.9219	0.0085	1.201
passt	1	0.07590	0.06863	1.2229	0.2688	1.079
nobreast	1	-0.12832	0.06047	4.5035	0.0338	0.880
lbw	1	-0.23642	0.08117	8.4832	0.0036	0.789
caesar	1	-0.07839	0.09272	0.7148	0.3979	0.925
multiple	1	-0.60731	0.21852	7.7240	0.0054	0.545
college	0	0	.	.	.	.

Why the differences? Well, like any fixed effects method, this one controls for all stable covariates, so it's possible that some of the earlier results in Table 10 were spurious. The thing to keep in mind is that, in this analysis, each woman is being compared to herself in a different birth interval. For a each woman, we're asking why some of her birth intervals are longer or shorter than others. Is it, for example, because she's married for some of the intervals and not for others? This approach will produce different answers than asking why some women tend to have longer birth intervals than others.

Despite the attractions of fixed effects Cox regression, it also has the usual disadvantages. As with other fixed effects methods we've employed, there may be a substantial loss of power compared with the conventional analysis. In this example, any woman with only one birth interval gets excluded because that interval can't be compared with any others. This eliminates 2109 birth intervals. Second, among women with exactly two birth intervals, if the second interval (which is always right censored) is shorter than the first, both intervals will be excluded. For the NSFG data, the elimination of these intervals results in the loss of another 1,468 cases.

Finally, even for those observations that are retained, the fixed effects method essentially discards information about variation *across* women and only uses variation *within* women. So if a particular covariate varies a great deal across women, but shows little variation over time for each woman, the coefficient for that variable will be poorly estimated. The variable PASST, for example, has 80 percent of its variance across women and only 20 percent within women. Not surprisingly, the standard error for its coefficient is more than twice as large in Table 11 as compared with Table 10, which was based on variation both within and between women.

CONCLUSION

By using each individual as his or her own control, fixed effects regression methods provide a relatively easy and effective way to control for stable variables that cannot be measured. For almost any kind of response variable, SAS has a procedure that can implement an appropriate fixed effects method. However, this paper does not cover all the possibilities that are discussed in my (2005) book. There, for example, I compare fixed effects with random effects models and show how fixed effects may be incorporated into PROC MIXED. This greatly expands the range of possible linear models. I also show how PROC CALIS can be used to estimate linear fixed effects models that embody reciprocal relationships among two or more variables. Finally, I show how PROC LOGISTIC may be used to estimate survival analysis models when events are not repeatable.

REFERENCES

Allison, P. D. (1995), *Survival Analysis Using SAS*, Cary, NC: SAS Institute Inc.
Allison, P. D. (1999), *Logistic Regression Using the SAS System*, Cary, NC: SAS Institute Inc.
Allison, P. D. (2002), "Bias in Fixed-Effects Cox Regression with Dummy Variables," unpublished paper, Department of Sociology, University of Pennsylvania.
Allison, P.D. (2005), *Fixed Effects Regression Methods for Longitudinal Data Using SAS,* Cary, NC: SAS Institute Inc.
Center for Human Resource Research (2002), *NLSY97 User's Guide*. Washington, DC: U.S. Department of Labor.
Cameron, A. C., and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge, UK: Cambridge University

Press.
Chamberlain, G. (1980), “Analysis of Covariance with Qualitative Data,” *Review of Economic Statistics*, 48, 225-238.
Greene, W. H. (2000), *Econometric Analysis*, 4th Edition, Upper Saddle River, NJ: Prentice Hall.
Kalbfleisch, J. D. and Sprott, D. A. (1970), “Applications of Likelihood Methods to Models Involving Large Numbers of Parameters” (with discussion), *Journal of the Royal Statistical Society, Series B*, 32, 175-208.
LaMotte, L. R. (1983), “Fixed-, Random-, and Mixed-Effects Models,” in *Encyclopedia of Statistical Sciences,* eds. S. Kotz, N. L. Johnson and C. B. Read, New York: John Wiley & Sons.
Sobel, M.E. (2000), “Causal Inference in the Social Sciences,” Journal of the American Statistical Association, 95, 647–651.
Wooldridge, J. M. (2001), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Paul D. Allison
University of Pennsylvania - Sociology
3718 Locust Walk
Philadelphia, PA 19104-6299
Work Phone: 215-898-6712
Fax: 215-573-2081
E-mail: allison@soc.upenn.edu
Web: http://www.ssc.upenn.edu/~allison

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies.