to symmetric distributions such as the normal and logistic, so that the probability can be conveniently written as Prob$(Y_{it} = y_{it}\,|\,\mathbf{x}_{it}) = P[q_{it}(\alpha_i + \mathbf{x}_{it}'\beta)]$. It will be convenient to let $z_{it} = \alpha_i + \mathbf{x}_{it}'\beta$ so Prob$(Y_{it} = y_{it}\,|\,\mathbf{x}_{it}) = P(q_{it}z_{it})$.

In our previous application of this model, in the linear regression case, we found that estimation of the parameters was made possible by a transformation of the data to deviations from group means which eliminated the person specific constants from the estimator. (See Section ~~9.4.1~~ 11.4.1) Save for the special case discussed later, that will not be possible here, so that if one desires to estimate the parameters of this model, it will be necessary actually to compute the possibly huge number of constant terms at the same time. This has been widely viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception. [See, e.g., Maddala (1987), p. 317.] The method for estimation of nonlinear fixed effects models such as the probit and logit models is detailed in Section ~~16.9.6.c~~ 14.9.6.d

The problems with the fixed effects estimator are statistical, not practical. The estimator relies on T_i increasing for the constant terms to be consistent—in essence, each α_i is estimated with T_i observations. But, in this setting, not only is T_i fixed, it is likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). The estimator of β is a function of the estimators of α, which means that the MLE of β is not consistent either. This is the incidental parameters problem. [See Neyman and Scott (1948) and Lancaster (2000).] There is, as well, a small sample (small T_i) bias in the estimators. How serious this bias is remains a question in the literature. Two pieces of received wisdom are Hsiao's (1986) results for a binary logit model [with additional results in Abrevaya (1997)] and Heckman and MaCurdy's (1980) results for the probit model. Hsiao found that for $T_i = 2$, the bias in the MLE of β is 100 percent, which is extremely pessimistic. Heckman and MaCurdy found in a Monte Carlo study that in samples of $n = 100$ and $T = 8$, the bias appeared to be on the order of 10 percent, which is substantive, but certainly less severe than Hsiao's results suggest. No other theoretical results have been shown for other models, although in *very* few cases, it can be shown that there is no incidental parameters problem. (The Poisson model mentioned in Chapter 16 14 is one of these special cases.) The fixed effects approach does have some appeal in that it does not require an assumption of orthogonality of the independent variables and the heterogeneity. An ongoing pursuit in the literature is concerned with the severity of the tradeoff of this virtue against the incidental parameters problem. Some commentary on this issue appears in Arellano (2001). Results of our own investigation appear in ~~Chapter 17~~ Section 15.5.2 and Greene (2004).

~~**Example 23.8 Binary Choice Models for Panel Data**~~

~~In Example 23.4 we fit a pooled binary logit model $y = 1(DocVis > 0)$ using the German health care utilization data examined in Example 11.11. The model is~~

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} + \beta_5 Education_{it} + \beta_6 Married_{it}).$$

~~No account of the panel nature of the data set was taken in that exercise. The sample contains a total of 27,326 observations on 7,293 families with T_i dispersed from one to seven. (See Example 11.11 for details.) Table 23.5 lists estimates of parameter estimates and estimated standard errors for probit and logit random and fixed effects models. There is a surprising amount of variation across the estimators. The coefficients are in bold to facilitate reading the table. It is generally difficult to compare across the estimators. The three estimators would~~

17-54

TABLE 23.6 Estimated Partial Effects for Panel Data Binary Choice Models

Model	*Age*	*Income*	*Kids*	*Education*	*Married*
Logit, P[a]	0.0048133	−0.043213	−0.053598	−0.010596	0.019936
Logit: RE,Q[b]	0.0064213	0.0035835	−0.035448	−0.010397	0.0041049
Logit: F,U[c]	0.024871	−0.014477	−0.020991	−0.027711	−0.013609
Logit: F,C[d]	0.0072991	−0.0043387	−0.0066967	−0.0078206	−0.0044842
Probit, P[a]	0.0048374	−0.043883	−0.053414	−0.010597	0.019783
Probit RE.Q[b]	0.0056049	−0.0008836	−0.042792	−0.0093756	0.0045426
Probit:RE,S[e]	0.0071455	−0.0010582	−0.054655	−0.011917	0.0059878
Probit: F,U[c]	0.023958	−0.013152	−0.018495	−0.027659	−0.012557

[a] Pooled estimator
[b] Butler and Moffitt estimator
[c] Unconditional fixed effects estimator
[d] Conditional fixed effects estimator
[e] Maximum simulated likelihood estimator

~~be expected to produce very different estimates in any of the three specifications—recall, for example, the pooled estimator is inconsistent in either the fixed or random effects cases. The logit results include two fixed effects estimators. The line market "U" is the unconditional (inconsistent) estimator. The one marked "C" is Chamberlain's consistent estimator. Note for all three fixed effects estimators, it is necessary to drop from the sample any groups that have *DocVis*$_{it}$ equal to zero or one for every period. There were 3,046 such groups, which is about 42 percent of the sample. We also computed the probit random effects model in two ways, first by using the Butler and Moffitt method, then by using maximum simulated likelihood estimation. In this case, the estimators are very similar, as might be expected. The estimated correlation coefficient, ρ, is computed as $\sigma_u^2/(\sigma_\varepsilon^2+\sigma_u^2)$. For the probit model, $\sigma_\varepsilon^2 = 1$. The MSL estimator computes $s_u = 0.9088376$, from which we obtained ρ. The estimated partial effects for the models are shown in Table 23.6. The average of the fixed effects constant terms is used to obtain a constant term for the fixed effects case. Once again there is a considerable amount of variation across the different estimators. On average, the fixed effects models tend to produce much larger values than the pooled or random effects models.~~

17.4.4 A Conditional Fixed Effects Estimator

FN 29

Why ~~did~~ does the incidental parameters problem arise here and not in the linear regression model?[29] Recall that estimation in the regression model was based on the deviations from group means, not the original data as it is here. The result we exploited there was that although $f(y_{it}\,|\,\mathbf{X}_i)$ is a function of α_i, $f(y_{it}\,|\,\mathbf{X}_i, \overline{y}_i)$ is not a function of α_i, and we used the latter in estimation of $\boldsymbol{\beta}$. In that setting, $\overline{y}_i$ is a **minimal sufficient statistic** for α_i. Sufficient statistics are available for a few distributions that we will examine, but not for the probit model. They are available for the logit model, as we now examine. KT

A fixed effects binary logit model is

$$\text{Prob}(y_{it} = 1\,|\,\mathbf{x}_{it}) = \frac{e^{\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}}}.$$

The unconditional likelihood for the nT independent observations is

$$L = \prod_i \prod_t (F_{it})^{y_{it}}(1 - F_{it})^{1-y_{it}}.$$

✓✓✓

KT Chamberlain (1980) [following Rasch (1960) and Andersen (1970)] observed that the **conditional likelihood function,**

$$L^c = \prod_{i=1}^{n} \text{Prob}\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i}\,\middle|\,\sum_{t=1}^{T_i} y_{it}\right),$$

FN 29 → on next page

17-55

✓ [29] The incidental parameters problem does show up in ML estimation of the FE linear model, where Neyman and Scott (1948) discovered it, in estimation of σ_ε^2. The MLE of σ_ε^2 is **e′e**/nT, which converges to $[(T-1)/T]\sigma_\varepsilon^2 < \sigma_\varepsilon^2$.

minus

17-56

is free of the incidental parameters, α_i. The joint likelihood for each set of T_i observations conditioned on the number of ones in the set is

$$\text{Prob}\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i} \,\middle|\, \sum_{t=1}^{T_i} y_{it}, \text{data}\right) = \frac{\exp\left(\sum_{t=1}^{T_i} y_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}{\sum_{\Sigma_t d_{it} = S_i} \exp\left(\sum_{t=1}^{T_i} d_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}.$$

~~46~~ 7 (17-4~~8~~)

The function in the denominator is summed over the set of all $\binom{T_i}{S_i}$ different sequences of T_i zeros and ones that have the same sum as $S_i = \sum_{t=1}^{T_i} y_{it}$.~~[25]~~30

FN 30

Consider the example of $T_i = 2$. The unconditional likelihood is

$$L = \prod_i \text{Prob}(Y_{i1} = y_{i1})\text{Prob}(Y_{i2} = y_{i2}).$$

For each pair of observations, we have these possibilities:

1. $y_{i1} = 0$ and $y_{i2} = 0$. $\text{Prob}(0, 0 \mid \text{sum} = 0) = 1$.
2. $y_{i1} = 1$ and $y_{i2} = 1$. $\text{Prob}(1, 1 \mid \text{sum} = 2) = 1$.

The ith term in L^c for either of these is just one, so they contribute nothing to the conditional likelihood function.~~[26]~~ 31 When we take logs, these terms (and these observations) will drop out. But suppose that $y_{i1} = 0$ and $y_{i2} = 1$. Then

FN 31

3. $\text{Prob}(0, 1 \mid \text{sum} = 1) = \dfrac{\text{Prob}(0, 1 \textit{ and } \text{sum} = 1)}{\text{Prob}(\text{sum} = 1)} = \dfrac{\text{Prob}(0, 1)}{\text{Prob}(0, 1) + \text{Prob}(1, 0)}$.

Therefore, for this pair of observations, the conditional probability is

$$\frac{\dfrac{1}{1 + e^{\alpha_i + \mathbf{x}_{i1}'\boldsymbol{\beta}}} \dfrac{e^{\alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta}}}}{\dfrac{1}{1 + e^{\alpha_i + \mathbf{x}_{i1}'\boldsymbol{\beta}}} \dfrac{e^{\alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta}}} + \dfrac{e^{\alpha_i + \mathbf{x}_{i1}'\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}_{i1}'\boldsymbol{\beta}}} \dfrac{1}{1 + e^{\alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta}}}} = \frac{e^{\mathbf{x}_{i2}'\boldsymbol{\beta}}}{e^{\mathbf{x}_{i1}'\boldsymbol{\beta}} + e^{\mathbf{x}_{i2}'\boldsymbol{\beta}}}.$$

By conditioning on the sum of the two observations, we have removed the heterogeneity. Therefore, we can construct the conditional likelihood function as the product of these terms for the pairs of observations for which the two observations are (0, 1). Pairs of observations with one and zero are included analogously. The product of the terms such as the preceding, for those observation sets for which the sum is not zero or T_i, constitutes the conditional likelihood. Maximization of the resulting function is straightforward and may be done by conventional methods.

As in the linear regression model, it is of some interest to test whether there is indeed heterogeneity. With homogeneity ($\alpha_i = \alpha$), there is no unusual problem, and the

30 ~~[25]~~The enumeration of all these computations stands to be quite a burden—see Arellano (2000, p. 47) or Baltagi (2005, p. 235). In fact, using a recursion suggested by Krailo and Pike (1984), the computation even with T_i up to 100 is routine.

31 ~~[26]~~Recall that in the probit model when we encountered this situation, the individual constant term could not be estimated and the group was removed from the sample. The same effect is at work here.

17-57

model can be estimated, as usual, as a logit model. It is not possible to test the hypothesis using the likelihood ratio test, however, because the two likelihoods are not comparable. (The conditional likelihood is based on a restricted data set.) None of the usual tests of restrictions can be used because the individual effects are never actually estimated.~~[27]~~[32] Hausman's (1978) specification test is a natural one to use here, however. Under the null hypothesis of homogeneity, both Chamberlain's conditional maximum likelihood estimator (CMLE) and the usual maximum likelihood estimator are consistent, but Chamberlain's is inefficient. (It fails to use the information that $\alpha_i = \alpha$, and it may not use all the data.) Under the alternative hypothesis, the unconditional maximum likelihood estimator is inconsistent,~~[28]~~[33] whereas Chamberlain's estimator is consistent and efficient. The Hausman test can be based on the chi-squared statistic

FN 32

FN 33

$$\chi^2 = (\hat{\beta}_{\text{CML}} - \hat{\beta}_{\text{ML}})'(\text{Var[CML]} - \text{Var[ML]})^{-1}(\hat{\beta}_{\text{CML}} - \hat{\beta}_{\text{ML}}). \qquad 17\text{-}48 \; \cancel{(23\text{-}42)}$$

47

The estimated covariance matrices are those computed for the two maximum likelihood estimators. For the unconditional maximum likelihood estimator, the row and column corresponding to the constant term are dropped. A large value will cast doubt on the hypothesis of homogeneity. (There are K degrees of freedom for the test.) It is possible that the covariance matrix for the maximum likelihood estimator will be larger than that for the conditional maximum likelihood estimator. If so, then the difference matrix in brackets is assumed to be a zero matrix, and the chi-squared statistic is therefore zero.

~~*Example 23.9 Fixed Effects Logit Models: Magazine Prices Revisited*~~

~~The fixed effects model does have some appeal, but the incidental parameters problem is a significant shortcoming of the unconditional probit and logit estimators. The conditional MLE for the fixed effects logit model is a fairly common approach. A widely cited application of the model is Cecchetti's (1986) analysis of changes in newsstand prices of magazines. Cecchetti's model was~~

$$\text{Prob}(\textit{Price change in year i of magazine t}) = \Lambda(\alpha_j + \mathbf{x}_{it}'\beta),$$

~~where the variables in $\mathbf{x}_{it}$ are (1) time since last price change, (2) inflation since last change, (3) previous fixed price change, (4) current inflation, (5) industry sales growth, and (6) sales volatility. The fixed effect in the model is indexed "*j*" rather than "*i*" as it is defined as a three-year interval for magazine *i*. Thus, a magazine that had been on the newstands for nine years would have three constants, not just one. In addition to estimating several specifications of the price change model, Cecchetti used the Hausman test in (23-42) to test for the existence of the common effects. Some of Cecchetti's results appear in Table 23.7.~~

~~Willis (2006) argued that Cecchetti's estimates were inconsistent and the Hausman test is invalid because right-hand-side variables (1), (2), and (6) are all functions of lagged dependent variables. This state dependence invalidates the use of the sum of the observations for the group as a sufficient statistic in the Chamberlain estimator and the Hausman tests. He proposes, instead, a method suggested by Heckman and Singer (1984b) to incorporate the unobserved heterogeneity in the *unconditional* likelihood function. The Heckman and Singer model can be formulated as a latent class model (see Sections 16.9.7 and 23.5.8) in which the classes are defined by different constant terms—the remaining parameters in the model are constrained to be equal across classes. Willis fit the Heckman and Singer model with~~

[32] ~~[27]~~ This produces a difficulty for this estimator that is shared by the semiparametric estimators discussed in the next section. Because the fixed effects are not estimated, it is not possible to compute probabilities or marginal effects with these estimated coefficients, and it is a bit ambiguous what one can do with the results of the computations. The brute force estimator that actually computes the individual effects might be preferable.

[33] ~~[28]~~ Hsiao (2003) derives the result explicitly for some particular cases.

~~to symmetric distributions such as the normal and logistic, so that the probability can be conveniently written as $\text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}) = P[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$. It will be convenient to let $z_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}$ so $\text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}) = P(q_{it} z_{it})$.~~

~~In our previous application of this model, in the linear regression case, we found that estimation of the parameters was made possible by a transformation of the data to deviations from group means which eliminated the person specific constants from the estimator. (See Section 9.4.1.) Save for the special case discussed later, that will not be possible here, so that if one desires to estimate the parameters of this model, it will be necessary actually to compute the possibly huge number of constant terms at the same time. This has been widely viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception. [See, e.g., Maddala (1987), p. 317.] The method for estimation of nonlinear fixed effects models such as the probit and logit models is detailed in Section 16.9.6.c.~~

~~The problems with the fixed effects estimator are statistical, not practical. The estimator relies on T_i increasing for the constant terms to be consistent—in essence, each α_i is estimated with T_i observations. But, in this setting, not only is T_i fixed, it is likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). The estimator of $\boldsymbol{\beta}$ is a function of the estimators of α, which means that the MLE of $\boldsymbol{\beta}$ is not consistent either. This is the incidental parameters problem. [See Neyman and Scott (1948) and Lancaster (2000).] There is, as well, a small sample (small T_i) bias in the estimators. How serious this bias is remains a question in the literature. Two pieces of received wisdom are Hsiao's (1986) results for a binary logit model [with additional results in Abrevaya (1997)] and Heckman and MaCurdy's (1980) results for the probit model. Hsiao found that for $T_i = 2$, the bias in the MLE of $\boldsymbol{\beta}$ is 100 percent, which is extremely pessimistic. Heckman and MaCurdy found in a Monte Carlo study that in samples of $n = 100$ and $T = 8$, the bias appeared to be on the order of 10 percent, which is substantive, but certainly less severe than Hsiao's results suggest. No other theoretical results have been shown for other models, although in *very* few cases, it can be shown that there is no incidental parameters problem. (The Poisson model mentioned in Chapter 16 is one of these special cases.) The fixed effects approach does have some appeal in that it does not require an assumption of orthogonality of the independent variables and the heterogeneity. An ongoing pursuit in the literature is concerned with the severity of the tradeoff of this virtue against the incidental parameters problem. Some commentary on this issue appears in Arellano (2001). Results of our own investigation appear in Chapter 17 and Greene (2004).~~

17.11 ***Example ~~23.6~~ Binary Choice Models for Panel Data***

In Example ~~23.4~~, we fit a pooled binary logit model $y = 1(DocVis > 0)$ using the German health care utilization data examined in ~~Example 11.11.~~ The model is 17.3 Appendix Table F7.1

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 + \beta_2\, Age_{it} + \beta_3\, Income_{it} + \beta_4\, Kids_{it} + \beta_5\, Education_{it} + \beta_6\, Married_{it}).$$

No account of the panel nature of the data set was taken in that exercise. The sample contains a total of 27,326 observations on 7,293 families with T_i dispersed from one to seven. (~~See Example 11.11 for details.~~) Table 23.5 lists estimates of parameter estimates and estimated standard errors for probit and logit random and fixed effects models. There is a surprising amount of variation across the estimators. The coefficients are in bold to facilitate reading the table. It is generally difficult to compare across the estimators. The three estimators would

17.5
17.8

17.8

17.8

TABLE ~~23.6~~ Estimated Parameters for Panel Data Binary Choice Models

			Variable					
Model	*Estimate*	*ln L*	*Constant*	*Age*	*Income*	*Kids*	*Education*	*Married*
Logit	β		0.25112	0.020709	−0.18592	−0.22947	−0.045587	0.085293
Pooled	St.Err.	−17673.10	0.091135	0.0012852	0.075064	0.029537	0.005646	0.033286
	Rob.SE[e]		0.12827	0.0017429	0.091546	0.038313	0.008075	0.045314
Logit R.E.	β	−15261.90	−0.13460	0.039267	0.021914	−0.21598	−0.063578	0.025071
$\rho = 0.41607$	St.Err.		0.17764	0.0024659	0.11866	0.047738	0.011322	0.056282
Logit	β	−9458.64		0.10475	−0.060973	−0.088407	−0.11671	−0.057318
F.E.(U)[a]	St.Err.			0.0072548	0.17829	0.074399	0.066749	0.10609
Logit	β	-6312.57 ~~−6299.02~~		0.084760	−0.050383	−0.077764	−0.090816	−0.052072
F.E.(C)[b]	St.Err.			0.0065022	0.15888	0.066282	0.056673	0.093044
Probit	β	−17670.94	0.15500	0.012835	−0.11643	−0.14118	−0.028115	0.052260
Pooled	St.Err.		0.056516	0.0007903	0.046329	0.018218	0.003503	0.020462
	Rob.SE[e]		0.079591	0.0010739	0.056543	0.023614	0.005014	0.027904
Probit:RE[c]	β	−16273.96	0.034113	0.020143	−0.003176	−0.15379	−0.033694	0.016325
$\rho = 0.44789$	St.Err.		0.096354	0.0013189	0.066672	0.027043	0.006289	0.031347
Probit:RE[d]	β	−16279.97	0.033290	0.020078	−0.002973	−0.153579	−0.033489	0.016826
$\rho = 0.44799$	St.Err.		0.063229	0.0009013	0.052012	0.020286	0.003931	0.022771
Probit	β	−9453.71		0.062528	−0.034328	−0.048270	−0.072189	−0.032774
F.E.(U)	St.Err.			0.0043219	0.10745	0.044559	0.040731	0.063627

[a] Unconditional fixed effects estimator
[b] Conditional fixed effects estimator
[c] Butler and Moffitt estimator
[d] Maximum simulated likelihood estimator
[e] Robust, "cluster" corrected standard error

.08384 −.06581 −.07802 −.12179 −.04847
(.006382) (.15743) (.066186) (.05466) (.092639)

17-60

TABLE ~~23.6~~ 17.9 Estimated Partial Effects for Panel Data Binary Choice Models

Model	*Age*	*Income*	*Kids*	*Education*	*Married*
Logit, P[a]	0.0048133	−0.043213	−0.053598	−0.010596	0.019936
Logit: RE,Q[b]	0.0064213	0.0035835	−0.035448	−0.010397	0.0041049
Logit: F,U[c]	0.024871	−0.014477	−0.020991	−0.027711	−0.013609
Logit: F,C[d]	0.0072991	−0.0043387	−0.0066967	−0.0078206	−0.0044842
Probit, P[a]	0.0048374	−0.043883	−0.053414	−0.010597	0.019783
Probit RE.Q[b]	0.0056049	−0.0008836	−0.042792	−0.0093756	0.0045426
Probit:RE,S[e]	0.0071455	−0.0010582	−0.054655	−0.011917	0.0059878
Probit: F,U[c]	0.023958	−0.013152	−0.018495	−0.027659	−0.012557

[a] Pooled estimator
[b] Butler and Moffitt estimator
[c] Unconditional fixed effects estimator
[d] Conditional fixed effects estimator
[e] Maximum simulated likelihood estimator

be expected to produce very different estimates in any of the three specifications—recall, for example, the pooled estimator is inconsistent in either the fixed or random effects cases. The logit results include two fixed effects estimators. The line market "U" is the unconditional (inconsistent) estimator. The one marked "C" is Chamberlain's consistent estimator. Note for all three fixed effects estimators, it is necessary to drop from the sample any groups that have *DocVis*$_{it}$ equal to zero or one for every period. There were 3,046 such groups, which is about 42 percent of the sample. We also computed the probit random effects model in two ways, first by using the Butler and Moffitt method, then by using maximum simulated likelihood estimation. In this case, the estimators are very similar, as might be expected. The estimated correlation coefficient, ρ, is computed as $\sigma_u^2/(\sigma_\varepsilon^2+\sigma_u^2)$. For the probit model, $\sigma_\varepsilon^2 = 1$. The MSL estimator computes $s_u = 0.9088376$, from which we obtained ρ. The estimated partial effects ~~for the models are shown in Table~~ ~~23.6~~ 17.9. The average of the fixed effects constant terms is used to obtain a constant term for the fixed effects case. Once again there is a considerable amount of variation across the different estimators. On average, the fixed effects models tend to produce much larger values than the pooled or random effects models.

T13 17.9 17.9

~~Why did the incidental parameters problem arise here and not in the linear regression model? Recall that estimation in the regression model was based on the deviations from group means, not the original data as it is here. The result we exploited there was that although $f(y_{it} \mid \mathbf{X}_i)$ is a function of α_i, $f(y_{it} \mid \mathbf{X}_i, \bar{y}_i)$ is not a function of α_i, and we used the latter in estimation of $\boldsymbol{\beta}$. In that setting, $\bar{y}_i$ is a **minimal sufficient statistic** for α_i. Sufficient statistics are available for a few distributions that we will examine, but not for the probit model. They are available for the logit model, as we now examine.~~

~~A fixed effects binary logit model is~~

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}) = \frac{e^{\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}}}.$$

~~The unconditional likelihood for the nT independent observations is~~

$$L = \prod_i \prod_t (F_{it})^{y_{it}} (1 - F_{it})^{1-y_{it}}.$$

~~Chamberlain (1980) [following Rasch (1960) and Andersen (1970)] observed that the **conditional likelihood function,**~~

$$L^c = \prod_{i=1}^{n} \text{Prob}\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i} \,\middle|\, \sum_{t=1}^{T_i} y_{it}\right),$$

17-61

~~model can be estimated, as usual, as a logit model. It is not possible to test the hypothesis using the likelihood ratio test, however, because the two likelihoods are not comparable. (The conditional likelihood is based on a restricted data set.) None of the usual tests of restrictions can be used because the individual effects are never actually estimated.[27] Hausman's (1978) specification test is a natural one to use here, however. Under the null hypothesis of homogeneity, both Chamberlain's conditional maximum likelihood estimator (CMLE) and the usual maximum likelihood estimator are consistent, but Chamberlain's is inefficient. (It fails to use the information that $\alpha_i = \alpha$, and it may not use all the data.) Under the alternative hypothesis, the unconditional maximum likelihood estimator is inconsistent,[28] whereas Chamberlain's estimator is consistent and efficient. The Hausman test can be based on the chi-squared statistic~~

$$\chi^2 = (\hat{\beta}_{\text{CML}} - \hat{\beta}_{\text{ML}})'(\text{Var}[\text{CML}] - \text{Var}[\text{ML}])^{-1}(\hat{\beta}_{\text{CML}} - \hat{\beta}_{\text{ML}}). \tag{23-42}$$

~~The estimated covariance matrices are those computed for the two maximum likelihood estimators. For the unconditional maximum likelihood estimator, the row and column corresponding to the constant term are dropped. A large value will cast doubt on the hypothesis of homogeneity. (There are K degrees of freedom for the test.) It is possible that the covariance matrix for the maximum likelihood estimator will be larger than that for the conditional maximum likelihood estimator. If so, then the difference matrix in brackets is assumed to be a zero matrix, and the chi-squared statistic is therefore zero.~~

12
17.8

Example ~~23.9~~ Fixed Effects Logit Models: Magazine Prices Revisited

The fixed effects model does have some appeal, but the incidental parameters problem is a significant shortcoming of the unconditional probit and logit estimators. The conditional MLE for the fixed effects logit model is a fairly common approach. A widely cited application of the model is Cecchetti's (1986) analysis of changes in newsstand prices of magazines. Cecchetti's model was

$$\text{Prob}(\textit{Price change in year i of magazine t}) = \Lambda(\alpha_j + \mathbf{x}_{it}'\beta),$$

TB 17.10

where the variables in $\mathbf{x}_{it}$ are (1) time since last price change, (2) inflation since last change, (3) previous fixed price change, (4) current inflation, (5) industry sales growth, and (6) sales volatility. The fixed effect in the model is indexed "j" rather than "i" as it is defined as a three-year interval for magazine i. Thus, a magazine that had been on the newstands for nine years would have three constants, not just one. In addition to estimating several specifications of the price change model, Cecchetti used the Hausman test in (~~23-42) to test for the existence~~ 47 17-48 of the common effects. Some of Cecchetti's results appear in Table ~~23.7.~~ 17.10

Willis (2006) argued that Cecchetti's estimates were inconsistent and the Hausman test is invalid because right-hand-side variables (1), (2), and (6) are all functions of lagged dependent variables. This state dependence invalidates the use of the sum of the observations for the group as a sufficient statistic in the Chamberlain estimator and the Hausman tests. He proposes, instead, a method suggested by Heckman and Singer (1984b) to incorporate the unobserved heterogeneity in the *unconditional* likelihood function. The Heckman and Singer model can be formulated as a latent class model (see Sections ~~16.9.7~~ 14.10 ~~and~~ ~~23.5.3~~ 17.4.7) in which the classes are defined by different constant terms—the remaining parameters in the model are constrained to be equal across classes. Willis fit the Heckman and Singer model with

~~[27] This produces a difficulty for this estimator that is shared by the semiparametric estimators discussed in the next section. Because the fixed effects are not estimated, it is not possible to compute probabilities or marginal effects with these estimated coefficients, and it is a bit ambiguous what one can do with the results of the computations. The brute force estimator that actually computes the individual effects might be preferable.~~

~~[28] Hsiao (2003) derives the result explicitly for some particular cases.~~

17.~~7~~10

TABLE ~~23.7~~ Models for Magazine Price Changes (standard errors in parentheses)

	Pooled	*Unconditional FE*	*Conditional FE Cecchetti*	*Conditional FE Willis*	*Heckman and Singer*
β_1	−1.10 (0.03)	−0.07 (0.03)	1.12 (3.66)	1.02 (0.28)	−0.09 (0.04)
β_2	6.93 (1.12)	8.83 (1.25)	11.57 (1.68)	19.20 (7.51)	8.23 (1.53)
β_5	−0.36 (0.98)	−1.14 (1.06)	5.85 (1.76)	7.60 (3.46)	−0.13 (1.14)
Constant 1	−1.90 (0.14)				−1.94 (0.20)
Constant 2					−29.15 (1.1e11)
ln *L*	−500.45	−473.18	−82.91	−83.72	−499.65
Sample size	1026	1026		543	1026

two classes to a restricted version of Cecchetti's model using variables (1), (2), and (5). The results in Table ~~23.7~~ 17.10 show some of the results from Willis's Table I. (Willis reports that he could not reproduce Cecchetti's results—the ones in Cecchetti's second column would be the counterparts—because of some missing values. In fact, Willis's estimates are quite far from Cecchetti's results, so it will be difficult to compare them. Both are reported here.) 17.10

The two "mass points" reported by Willis are shown in Table ~~23.7~~. He reports that these two values (−1.94 and −29.15) correspond to class probabilities of 0.88 and 0.12, though it is difficult to make the translation based on the reported values. He does note that the change in the log-likelihood in going from one mass point (pooled logit model) to two is marginal, only from −500.45 to −499.65. There is another anomaly in the results that is consistent with this finding. The reported standard error for the second "mass point" is 1.1×10^{11}, or essentially $+\infty$. The finding is consistent with overfitting the latent class model. The results suggest that the better model is a one-class (pooled) model.

~~23.5.3 MODELING HETEROGENEITY~~

~~The panel data analysis considered thus far has focused on modeling heterogeneity with the fixed and random effects specifications. Both assume that the heterogeneity is continuously distributed among individuals. The random effects model is fully parametric, requiring a full specification of the likelihood for estimation. The fixed effects model is essentially semiparametric. It requires no specific distributional assumption, however, it does require that the realizations of the latent heterogeneity be treated as parameters, either estimated in the unconditional fixed effects estimator or conditioned out of the likelihood function when possible. As noted in the preceding example, Heckman and Singer's (1984b) model provides a less stringent model specification based on a discrete distribution of the latent heterogeneity. A straightforward method of implementing their model is to cast it as a latent class model in which the classes are distinguished by different constant terms and the associated probabilities. The class probabilities are treated as parameters to be estimated with the model parameters.~~

~~*Example 23.10 Semiparametric Models of Heterogeneity*~~

~~We have extended the random effects and fixed effects logit models in Example 23.8 by fitting the Heckman and Singer (1984b) model. Table 23.8 shows the specification search and the results under different specifications. The first column of results shows the estimated fixed effects model from Example 23.8. The conditional estimates are shown in parentheses. Of the 7,293 groups in the sample, 3,056 are not used in estimation of the fixed effects models because the sum of Doctor$_{it}$ is either 0 or T$_i$ for the group. The mean and standard deviation of the estimated underlying heterogeneity distribution are computed using the estimates of~~

17.4.5 Mundlak's Approach, Variable Addition and Bias Reduction

Thus far, both the fixed effects (FE) and the random effects (RE) specifications present problems for modeling binary choice with panel data. The MLE of the FE model is inconsistent even when the model is properly specified – this is the incidental parameters problem. (And, like the linear model, the FE probit and logit models do not allow time invariant regressors.) The random effects specification requires a strong, often unreasonable assumption that the effects and the regressors are uncorrelated. Of the two, the FE model is the more appealing, though with modern longitudinal data sets with many demographics, the problem of time invariant variables would seem to be compelling. This would seem to recommend the conditional estimator in Section 17.4.4, save for yet another complication. With no estimates of the constant terms, neither probabilities nor partial effects can be computed with the results. We are left making inferences about ratios of coefficient. Two approaches have been suggested for finding a middle ground: Mundlak's (1978) approach that involves projecting the effects on the group means of the time varying variables and recent developments such as Fernandez-Val's approach that involves correcting the bias in the FE MLE.

The Mundlak (1978) [and Chamberlain (1984) and Wooldridge, e.g., (2002a)] approach augments (17-45) as follows: =44

$$
\begin{aligned}
y_{it}^* &= \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}\\
\text{Prob}(y_{it} = 1|\mathbf{x}_{it}) &= F(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})\\
\alpha_i &= \alpha + \bar{\mathbf{x}}_i'\boldsymbol{\delta} + u_i,
\end{aligned}
$$

where we have used $\bar{\mathbf{x}}_i$ generically for the group means of the time varying variables in $\mathbf{x}_{it}$. The reduced form of the model is

$$\text{Prob}(y_{it} = 1|\mathbf{x}_{it}) = \text{F}(\alpha + \bar{\mathbf{x}}_i'\boldsymbol{\delta} + \mathbf{x}_{it}'\boldsymbol{\beta} + u_i).$$

(Wooldridge and Chamberlain also suggest using all years of $\mathbf{x}_{it}$ rather than the group means. This raises a problem in unbalanced panels, however. We will ignore this possibility.) The projection of α_i on $\bar{\mathbf{x}}_i$ produces a random effects formulation. As in the linear model (see Section 11.5.6), it also suggests a means of testing for fixed vs. random effects. Since $\boldsymbol{\delta} = \mathbf{0}$ produces the pure random effects model, a joint Wald test of the null hypothesis that $\boldsymbol{\delta}$ equals zero can be used.

Example 17.13 Panel Data Random Effects Estimators

Example 17.11 presents several estimators of panel data estimators for the probit and logit models. Pooled, random effects and fixed effects estimates are given for the probit model

$$\text{Prob}(\textit{DocVis}_{it} > 0) = \Phi(\beta_1 + \beta_2\,\textit{Age}_{it} + \beta_3\,\textit{Income}_{it} + \beta_4\,\textit{Kids}_{it} + \beta_5\,\textit{Education}_{it} + \beta_6\,\textit{Married}_{it}).$$

TB 17.11

We continue that analysis here by considering Mundlak's approach to the common effects model. Table 17.11 presents the random effects model from earlier, and the augmented estimator that contains the group means of the variables, all of which are time varying. The addition of the group means to the regression brings large changes to the estimates of the parameters, which might suggest the appropriateness of the fixed effects model. A formal test is carried by computing a Wald statistic for the null hypothesis that the last five coefficients in the augmented model equal zero. The chi squared statistic equals 113.282 with five degrees of freedom. The critical value from the chi squared table for 95% significance is =50

(percent)

minus

11.07, so the hypothesis that δ equals zero, that is, the hypothesis of the random effects model (restrictions), is rejected. The two log likelihoods are -16273.96 for the REM and -16222.06 for the augmented REM. The LR statistic would be twice the difference, or 103.8. This produces the same conclusion. The FEM appears to be the preferred model.

minus

Table 17.11 Estimated Random Effects Models

	Constant	Age	Income	Kids	Education	Married
Random Effects	0.03411 (0.09635)	0.02014 (0.00132)	-0.00318 (0.06667)	-0.15379 (0.02704)	-0.03369 (0.00629)	0.01633 (0.03135)
Augmented Model	0.37485 (0.10501)	0.05035 (0.00357)	-0.03057 (0.09318)	-0.04202 (0.03751)	-0.05449 (0.03307)	-0.02645 (0.05180)
Means		-0.03659 (0.00384)	-0.35065 (0.13984)	-0.22509 (0.05499)	0.02387 (0.03374)	0.14668 (0.06607)

Au: Sub "FEMALE" ok italic?

has

A series of recent studies have sought to maintain the fixed effects specification while correcting the bias due to the incidental parameters problem. There are two broad approaches. Hahn and Kuersteiner (2004), Hahn and Newey (2005) and Fernandez-Val (2009) have developed an approximate, "large *T*" result for plim$\left(\hat{\beta}_{FE,MLE} - \beta\right)$ that produces a direct correction to the estimator, itself. Fernandez-Val (2009) develops corrections for the estimated constant terms as well. Arellano and Hahn (2006, 2007) propose a modification of the log likelihood function with, in turn, different first order estimation equations, that produces an approximately unbiased estimator of **β**. In a similar fashion to the second of these approaches, Carro (2007) modifies the first order conditions (estimating equations) from the original log likelihood function, once again to produce an approximately unbiased estimator of **β**. (In general, given the overall approach of using a large *T* approximation, the payoff to these estimators is to reduce the bias of the FE-MLE from O(1/*T*) to O(1/T^2), which is a considerable reduction.) These estimators are not yet in widespread use. The received evidence suggests that in the simple case we are considering here, the incidental parameters problem is a secondary concern when *T* reaches say 10 or so. For some modern public use data sets, such as the BHPS or GSOEP which are beyond their 15th wave, the incidental parameters problem may not be too severe. However, most of the studies mentioned ~~above~~ are concerned with dynamic models (see section 17.4.6 following), where the problem is possible more severe than in the static case. Research in this area is ongoing.

FEMALE

FEMALE

~~where $w_i = y_i(\omega_1/p_1) + (1 - y_i)(\omega_0/p_0)$. Note that w_i takes only two different values. The derivatives and the Hessian are likewise weighted. A final correction is needed after estimation; the appropriate estimator of the asymptotic covariance matrix is the sandwich estimator discussed in Section 23.4.1, $\mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}$ (with weighted $\mathbf{B}$ and $\mathbf{H}$), instead of $\mathbf{B}$ or $\mathbf{H}$ alone. (The weights are not squared in computing $\mathbf{B}$.)[20]~~

17.4.6 ~~23.4.7~~ DYNAMIC BINARY CHOICE MODELS

A random or fixed effects model that explicitly allows for lagged effects would be

$$y_{it} = \mathbf{1}(\mathbf{x}_{it}'\beta + \alpha_i + \gamma y_{i,t-1} + \varepsilon_{it} > 0).$$

Lagged effects, or **persistence,** in a binary choice setting can arise from three sources, serial correlation in ε_{it}, the **heterogeneity,** α_i, or **true state dependence** through the term $\gamma y_{i,t-1}$. Chiappori (1998) [and see Arellano (2001)] suggests an application to the French automobile insurance market in which the incentives built into the pricing system are such that having an accident in one period should lower the probability of having one in the next (state dependence), but, some drivers remain more likely to have accidents than others in every period, which would reflect the heterogeneity instead. State dependence is likely to be particularly important in the typical panel which has only a few observations for each individual. Heckman (1981a) examined this issue at length. Among his findings were that the somewhat muted small sample bias in fixed effects models with $T = 8$ was made much worse when there was state dependence. A related problem is that with a relatively short panel, the **initial conditions,** y_{i0}, have a crucial impact on the entire path of outcomes. Modeling dynamic effects and initial conditions in binary choice models is more complex than in the linear model, and by comparison there are relatively fewer firm results in the applied literature.~~[21]~~ [34]

FN 34

~~Much of the contemporary literature has focused on methods of avoiding the strong parametric assumptions of the probit and logit models. Manski (1987) and Honore and Kyriazidou (2000) show that Manski's (1986) maximum score estimator can be applied to the differences of unequal pairs of observations in a two period panel with fixed effects. However, the limitations of the maximum score estimator have motivated research on other approaches. An extension of lagged effects to a parametric model is Chamberlain (1985), Jones and Landwehr (1988), and Magnac (1997), who added state dependence to Chamberlain's fixed effects logit estimator. Unfortunately, once the identification issues are settled, the model is only operational if there are no other exogenous variables in it, which limits is usefulness for practical application. Lewbel (2000) has extended his fixed effects estimator to dynamic models as well. In this framework, the narrow assumptions about the independent variables somewhat~~

~~[20] WESML and the choice-based sampling estimator are not the free lunch they may appear to be. That which the biased sampling does, the weighting undoes. It is common for the end result to be very large standard errors, which might be viewed as unfortunate, insofar as the purpose of the biased sampling was to balance the data precisely to avoid this problem.~~

[34] ~~[21]~~ A survey of some of these results is given by Hsiao (2003). Most of Hsiao (2003) is devoted to the linear regression model. A number of studies specifically focused on discrete choice models and panel data have appeared recently, including Beck, Epstein, Jackman, and O'Halloran (2001), Arellano (2001), and Greene (2001). Vella and Verbeek (1998) provide an application to the joint determination of wages and union membership. Other important references are Aguirregabiria and Mira (2010), Carro (2007), and Fernandez-Val (2009). Stewart (2006) and Arulampalam and Stewart (2007) provide several results for practitioners.

The correlation between α_i and $y_{i,t-1}$ in the dynamic binary choice model makes $y_{i,t-1}$ endogenous. Thus, the estimators we have examined thus far will not be consistent. Two familiar alternative approaches that have appeared in recent applications are due to Heckman (1981) and Wooldridge (2005), both of which build on the random effects specification. Heckman's approach provides a separate equation for the initial condition,

$$\text{Prob}(y_{i1} = 1|\mathbf{x}_{i1},\mathbf{z}_i,\alpha_i) = \Phi(\mathbf{x}_{i1}'\boldsymbol{\delta} + \mathbf{z}_i'\boldsymbol{\tau} + \theta\alpha_i)$$

$$\text{Prob}(y_{it} = 1|\mathbf{x}_{it},y_{i,t-1},\alpha_i) = \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i),\ t = 2,\ldots,T_i,$$

where $\mathbf{z}_i$ is a set of "instruments" observed at the first period that are not contained in $\mathbf{x}_{it}$. The conditional log likelihood is

$$\begin{aligned}\ln L|\boldsymbol{\alpha} &= \sum_{i=1}^{n} \ln\left\{\Phi[(2y_{i1}-1)(\mathbf{x}_{i1}'\boldsymbol{\delta} + \mathbf{z}_i'\boldsymbol{\tau} + \theta\alpha_i)]\prod_{t=2}^{T_i}\Phi\left[(2y_{it}-1)(\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i)\right]\right\} \\ &= \sum_{i=1}^{n} \ln L_i \,|\, \alpha_i.\end{aligned}$$

We now adopt the random effects approach and further assume that α_i is normally distributed with mean zero and variance σ_α^2. The random effects log likelihood function can be maximized with respect to $(\boldsymbol{\delta},\boldsymbol{\tau},\theta,\boldsymbol{\beta},\gamma,\sigma_\alpha)$ using either the Butler and Moffitt quadrature method or the maximum simulated likelihood method described in Section 17.4.2. Stewart and Arulampalam (2007) suggest a useful shortcut for formulating the Heckman model. Let $D_{it} = 1$ in period 1 and 0 in every other period and let $C_{it} = 1 - D_{it}$. Then, the two parts may be combined in

$$\ln L|\boldsymbol{\alpha} = \sum_{i=1}^{n} \ln\prod_{t=1}^{T_i}\left\{\Phi\left[(2y_{it}-1)\left\langle C_{it}(\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i,t-1}) + D_{it}(\mathbf{x}_{it}'\boldsymbol{\delta} + \mathbf{z}_i'\boldsymbol{\tau}) + (1+\lambda D_{it})\alpha_i\right\rangle\right]\right\}.$$

In this form, the model can be viewed as a random parameters (random constant term) model in which there is heteroscedasticity in the random part of the constant term.

Wooldridge's approach builds on the Mundlak device of the previous section. Starting from the same point, he suggests a model for the random effect conditioned on the initial value. Thus,

$$\alpha_i|y_{i1},\mathbf{z}_i \sim N[\alpha_0 + \eta y_{i1} + \mathbf{z}_i'\boldsymbol{\tau},\ \sigma_\alpha^2].$$

Assembling the parts, Wooldridge's model is a bit simpler than Heckman's;

$$\text{Prob}(y_{it} = 1|\mathbf{x}_{it},y_{i1},u_i) = \Phi[(2y_{it}-1)(\alpha_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \eta y_{i1} + \mathbf{z}_i'\boldsymbol{\tau} + u_i)],\ \text{t} = 2,\ldots,T_i.$$

where $w_i = y_i(\omega_1/p_1) + (1 - y_i)(\omega_0/p_0)$. Note that w_i takes only two different values. The derivatives and the Hessian are likewise weighted. A final correction is needed after estimation; the appropriate estimator of the asymptotic covariance matrix is the sandwich estimator discussed in Section 23.4.1, $\mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}$ (with weighted **B** and **H**), instead of **B** or **H** alone. (The weights are not squared in computing **B**.)[20]

23.4.7 DYNAMIC BINARY CHOICE MODELS

A random or fixed effects model that explicitly allows for lagged effects would be

$$y_{it} = \mathbf{1}(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma y_{i,t-1} + \varepsilon_{it} > 0).$$

Lagged effects, or **persistence,** in a binary choice setting can arise from three sources, serial correlation in ε_{it}, the **heterogeneity,** α_i, or true **state dependence** through the term $\gamma y_{i,t-1}$. Chiappori (1998) [and see Arellano (2001)] suggests an application to the French automobile insurance market in which the incentives built into the pricing system are such that having an accident in one period should lower the probability of having one in the next (state dependence), but, some drivers remain more likely to have accidents than others in every period, which would reflect the heterogeneity instead. State dependence is likely to be particularly important in the typical panel which has only a few observations for each individual. Heckman (1981a) examined this issue at length. Among his findings were that the somewhat muted small sample bias in fixed effects models with $T = 8$ was made much worse when there was state dependence. A related problem is that with a relatively short panel, the **initial conditions,** y_{i0}, have a crucial impact on the entire path of outcomes. Modeling dynamic effects and initial conditions in binary choice models is more complex than in the linear model, and by comparison there are relatively fewer firm results in the applied literature.[21]

Much of the contemporary literature has focused on methods of avoiding the strong parametric assumptions of the probit and logit models. Manski (1987) and Honore and Kyriazidou (2000) show that Manski's (1986) maximum score estimator can be applied to the differences of unequal pairs of observations in a two period panel with fixed effects. However, the limitations of the maximum score estimator have motivated research on other approaches. An extension of lagged effects to a parametric model is Chamberlain (1985), Jones and Landwehr (1988), and Magnac (1997), who added state dependence to Chamberlain's fixed effects logit estimator. Unfortunately, once the identification issues are settled, the model is only operational if there are no other exogenous variables in it, which limits is usefulness for practical application. Lewbel (2000) has extended his fixed effects estimator to dynamic models as well. In ~~this framework, the narrow assumptions about the independent variables somewhat~~

its

[20] WESML and the choice-based sampling estimator are not the free lunch they may appear to be. That which the biased sampling does, the weighting undoes. It is common for the end result to be very large standard errors, which might be viewed as unfortunate, insofar as the purpose of the biased sampling was to balance the data precisely to avoid this problem.

[21] A survey of some of these results is given by Hsiao (2003). Most of Hsiao (2003) is devoted to the linear regression model. A number of studies specifically focused on discrete choice models and panel data have appeared recently, including Beck, Epstein, Jackman and O'Halloran (2001), Arellano (2001) and Greene (2001). Vella and Verbeek (1998) provide an application to the joint determination of wages and union membership.

AU: Run-in OK? Or new ¶?

17-68

Dong and Lewbel (2010) have extended Lewbel's "special regressor" method to dynamic binary choice models and have devised an estimator based on an IV linear regression.

~~limit its practical applicability.~~ Honore and Kyriazidou (2000) have combined the logic of the **conditional logit model** and Manski's maximum score estimator. They specify

KT

$$\text{Prob}(y_{i0} = 1 \mid \mathbf{x}_i, \alpha_i) = p_0(\mathbf{x}_i, \alpha_i) \quad \text{where } \mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}),$$

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_i, \alpha_i, y_{i0}, y_{i1}, \ldots, y_{i,t-1}) = F(\mathbf{x}_{it}'\beta + \alpha_i + \gamma y_{i,t-1}) \quad t = 1, \ldots, T.$$

The analysis assumes a single regressor and focuses on the case of $T = 3$. The resulting estimator resembles Chamberlain's but relies on observations for which $\mathbf{x}_{it} = \mathbf{x}_{i,t-1}$, which rules out direct time effects as well as, for practical purposes, any continuous variable. The restriction to a single regressor limits the generality of the technique as well. The need for observations with equal values of $\mathbf{x}_{it}$ is a considerable restriction, and the authors propose a kernel density estimator for the difference, $\mathbf{x}_{it} - \mathbf{x}_{i,t-1}$, instead which does relax that restriction a bit. The end result is an estimator that converges (they conjecture) but to a nonnormal distribution and at a rate slower than $n^{-1/3}$.

Semiparametric estimators for dynamic models at this point in the development are still primarily of theoretical interest. Models that extend the parametric formulations to include state dependence have a much longer history, including Heckman (1978, 1981a, 1981b), Heckman and MaCurdy (1980), Jakubson (1988), Keane (1993), and Beck et al. (2001) to name a few.[~~27~~ 35] In general, even without heterogeneity, dynamic models ultimately involve modeling the joint outcome $(y_{i0}, \ldots, y_{iT})$, which necessitates some treatment involving multivariate integration. Example ~~23.7~~ ~~17.15~~ 17.14 describes ~~a recent~~ an application. Stewart (2006) provides another.

FN 35

Example ~~23.7~~ 17.14 An Intertemporal Labor Force Participation Equation

Hyslop (1999) presents a model of the labor force participation of married women. The focus of the study is the high degree of persistence in the participation decision. Data used in the study were the years 1979–1985 of the Panel Study of Income Dynamics. A sample of 1,812 continuously married couples were studied. Exogenous variables that appeared in the model were measures of permanent and transitory income and fertility captured in yearly counts of the number of children from 0–2, 3–5, and 6–17 years old. Hyslop's formulation, in general terms, is

$$\text{(initial condition) } y_{i0} = 1(\mathbf{x}_{i0}'\beta_0 + v_{i0} > 0),$$
$$\text{(dynamic model) } y_{it} = 1(\mathbf{x}_{it}'\beta + \gamma y_{i,t-1} + \alpha_i + v_{it} > 0)$$
$$\text{(heterogeneity correlated with participation) } \alpha_i = \mathbf{z}_i'\delta + \eta_i,$$
$$\text{(stochastic specification)}$$

$$\eta_i \mid \mathbf{X}_i \sim N[0, \sigma_\eta^2],$$
$$v_{i0} \mid \mathbf{X}_i \sim N[0, \sigma_0^2],$$
$$w_{it} \mid \mathbf{X}_i \sim N[0, \sigma_w^2],$$
$$v_{it} = \rho v_{i,t-1} + w_{it}, \sigma_\eta^2 + \sigma_w^2 = 1.$$
$$\text{Corr}[v_{i0}, v_{it}] = \rho^t, \quad t = 1, \ldots, T-1.$$

[~~27~~ 35] Beck et al. (2001) is a bit different from the others mentioned in that in their study of "state failure," they observe a large sample of countries (147) observed over a fairly large number of years, 40. As such, they are able to formulate their models in a way that makes the asymptotics with respect to T appropriate. They can analyze the data essentially in a time-series framework. Sepanski (2000) is another application that combines state dependence and the random coefficient specification of Akin, Guilkey, and Sickles (1979).

17-69

The presence of the autocorrelation and state dependence in the model invalidate the simple maximum likelihood procedures we examined earlier. The appropriate likelihood function is constructed by formulating the probabilities as

$$\text{Prob}(y_{i0}, y_{i1}, \ldots) = \text{Prob}(y_{i0}) \times \text{Prob}(y_{i1} \mid y_{i0}) \times \cdots \times \text{Prob}(y_{iT} \mid y_{i,T-1}).$$

This still involves a $T = 7$ order normal integration, which is approximated in the study using a simulator similar to the GHK simulator discussed in 17.3.3 15.6.2.b Among Hyslop's results are a comparison of the model fit by the simulator for the multivariate normal probabilities with the same model fit using the maximum simulated likelihood technique described in Section ~~17.5.1.~~ 15.6

23.5 BINARY CHOICE MODELS FOR PANEL DATA

Qualitative response models have been a growth industry in econometrics. The recent literature, particularly in the area of panel data analysis, has produced a number of new techniques. The availability of high-quality panel data sets on microeconomic behavior has maintained an interest in extending the models of Chapter 9 to binary (and other discrete choice) models. In this section, we will survey a few results from this rapidly growing literature.

The structural model for a possibly unbalanced panel of data would be written

$$y^*_{it} = \mathbf{x}'_{it}\beta + \varepsilon_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T_i,$$
$$y_{it} = 1 \quad \text{if } y^*_{it} > 0, \text{ and 0 otherwise.}$$

The second line of this definition is often written

$$y_{it} = \mathbf{1}(\mathbf{x}'_{it}\beta + \varepsilon_{it} > 0)$$

to indicate a variable that equals one when the condition in parentheses is true and zero when it is not. Ideally, we would like to specify that ε_{it} and ε_{is} are freely correlated within a group, but uncorrelated across groups. But doing so will involve computing joint probabilities from a T_i variate distribution, which is generally problematic.[23] (We will return to this issue later.) A more promising approach is an effects model,

$$y^*_{it} = \mathbf{x}'_{it}\beta + v_{it} + u_i, \quad i = 1, \ldots, n, t = 1, \ldots, T_i,$$
$$y_{it} = 1 \quad \text{if } y^*_{it} > 0, \text{ and 0 otherwise,}$$

where, as before (see Section 9.5), u_i is the unobserved, individual specific heterogeneity. Once again, we distinguish between "random" and "fixed" effects models by the relationship between u_i and $\mathbf{x}_{it}$. The assumption that u_i is unrelated to $\mathbf{x}_{it}$, so that the conditional distribution $f(u_i \mid \mathbf{x}_{it})$ is not dependent on $\mathbf{x}_{it}$, produces the **random effects model.** Note that this places a restriction on the distribution of the heterogeneity.

[23] A "limited information" approach based on the GMM estimation method has been suggested by Avery, Hansen, and Hotz (1983). With recent advances in simulation-based computation of multinormal integrals (see Section 17.5.1), some work on such a panel data estimator has appeared in the literature. See, for example, Geweke, Keane, and Runkle (1994, 1997). The GEE estimator of Diggle, Liang, and Zeger (1994) [see also, Liang and Zeger (1986) and Stata (2006)] seems to be another possibility. However, in all these cases, it must be remembered that the procedure specifies estimation of a correlation matrix for a T_i vector of unobserved variables based on a dependent variable that takes only two values. We should not be too optimistic about this if T_i is even moderately large.

17-70

TABLE 23.7 Models for Magazine Price Changes (standard errors in parentheses)

	Pooled	*Unconditional FE*	*Conditional FE Cecchetti*	*Conditional FE Willis*	*Heckman and Singer*
β_1	−1.10 (0.03)	−0.07 (0.03)	1.12 (3.66)	1.02 (0.28)	−0.09 (0.04)
β_2	6.93 (1.12)	8.83 (1.25)	11.57 (1.68)	19.20 (7.51)	8.23 (1.53)
β_5	−0.36 (0.98)	−1.14 (1.06)	5.85 (1.76)	7.60 (3.46)	−0.13 (1.14)
Constant 1	−1.90 (0.14)				−1.94 (0.20)
Constant 2					−29.15 (1.1e11)
ln *L*	−500.45	−473.18	−82.91	−83.72	−499.65
Sample size	1026	1026		543	1026

two classes to a restricted version of Cecchetti's model using variables (1), (2), and (5). The results in Table 23.7 show some of the results from Willis's Table I. (Willis reports that he could not reproduce Cecchetti's results—the ones in Cecchetti's second column would be the counterparts—because of some missing values. In fact, Willis's estimates are quite far from Cecchetti's results, so it will be difficult to compare them. Both are reported here.)

The two "mass points" reported by Willis are shown in Table 23.7. He reports that these two values (−1.94 and −29.15) correspond to class probabilities of 0.88 and 0.12, though it is difficult to make the translation based on the reported values. He does note that the change in the log-likelihood in going from one mass point (pooled logit model) to two is marginal, only from −500.45 to −499.65. There is another anomaly in the results that is consistent with this finding. The reported standard error for the second "mass point" is 1.1×10^{11}, or essentially $+\infty$. The finding is consistent with overfitting the latent class model. The results suggest that the better model is a one-class (pooled) model.

17.4.~~8~~7 ~~23.6.3~~ A SEMIPARAMETRIC MODEL FOR INDIVIDUAL ~~MODELING~~ HETEROGENEITY

The panel data analysis considered thus far has focused on modeling heterogeneity with the fixed and random effects specifications. Both assume that the heterogeneity is continuously distributed among individuals. The random effects model is fully parametric, requiring a full specification of the likelihood for estimation. The fixed effects model is essentially semiparametric. It requires no specific distributional assumption; however, it does require that the realizations of the latent heterogeneity be treated as parameters, either estimated in the unconditional fixed effects estimator or conditioned out of the likelihood function when possible. As noted in the preceding example, Heckman and Singer's (1984b) model provides a less stringent model specification based on a discrete distribution of the latent heterogeneity. A straightforward method of implementing their model is to cast it as a latent class model in which the classes are distinguished by different constant terms and the associated probabilities. The class probabilities are treated as parameters to be estimated with the model parameters.

Example ~~23.10~~ 17.15 ***Semiparametric Models of Heterogeneity*** ~~17.13~~ 17.12 17.11

We have extended the random effects and fixed effects logit models in Example ~~23.8~~ by fitting the Heckman and Singer (1984b) model. Table ~~23.8~~ shows the specification search and the results under different specifications. The first column of results shows the estimated fixed effects model from Example ~~23.8~~. The conditional estimates are shown in parentheses. Of the 7,293 groups in the sample, 3,056 are not used in estimation of the fixed effects models because the sum of Doctor_{it} is either 0 or T_i for the group. The mean and standard deviation of the estimated underlying heterogeneity distribution are computed using the estimates of

TB 17.12 17.11

17.12

TABLE ~~23.8~~ Estimated Heterogeneity Models

		Number of Classes				
	Fixed Effect	*1*	*2*	*3*	*4*	*5*
β_1	0.10475 (0.084760)	0.020708	0.030325	0.033684	0.034083	0.034159
β_2	−0.060973 (−0.050383)	−0.18592	0.025550	−0.0058013	−0.0063516	−0.013627
β_3	−0.088407 (−0.077764)	−0.22947	−0.24708	−0.26388	−0.26590	−0.26626
β_4	−0.11671 (−0.090816)	−0.045588	−0.050924	−0.058022	−0.059751	−0.059176
β_5	−0.057318 (−0.52072)	0.085293	0.042974	0.037944	0.029227	0.030699
α_1	−2.62334	0.25111 (1.00000)	0.91764 (0.62681)	1.71669 (0.34838)	1.94536 (0.29309)	2.76670 (0.11633)
α_2			−1.47800 (0.37319)	−2.23491 (0.18412)	−1.76371 (0.21714)	1.18323 (0.26468)
α_3				−0.28133 (0.46749)	−0.036739 (0.46341)	−1.96750 (0.19573)
α_4					−4.03970 (0.026360)	−0.25588 (0.40930)
α_5						−6.48191 (0.013960)
Mean	−2.62334	0.00000	0.023613	0.055059	0.063685	0.054705
Std. Dev.	3.13415	0.00000	1.158655	1.40723	1.48707	1.62143
ln *L*	−9458.638 (−6299.02)	−17673.10	−16353.14	−16278.56	−16276.07	−16275.85
AIC	1.00349	1.29394	1.19748	1.19217	1.19213	1.19226

α_i for the remaining 4,237 groups. The remaining five columns in the table show the results for different numbers of latent classes in the Heckman and Singer model. The listed constant terms are the "mass points" of the underlying distributions. The associated class probabilities are shown in parentheses under them. The mean and standard deviation are derived from the 2- to 5-point discrete distributions shown. It is noteworthy that the mean of the distribution is relatively stable, but the standard deviation rises monotonically. The search for the best model would be based on the AIC. As noted in Section 16.9.6, using a likelihood ratio test in this context is dubious, as the number of degrees of freedom is ambiguous. Based on the AIC, the four-class model is the preferred specification. 14 16.10

MODELING

17.4.8 ~~23.5.4~~ PARAMETER HETEROGENEITY

Section 11.11

In ~~Chapter 9~~, we examined specifications that extend the underlying heterogeneity to all the parameters of the model. We have considered two approaches. The random parameters, or mixed models discussed in Chapter ~~17~~ allow parameters to be distributed continuously across individuals. The latent class model in Section 16.9.6 specifies a discrete distribution instead. (The Heckman and Singer model in the previous section applies this method to the constant term.) Most of the focus to this point, save for Example 16.16, has been on linear models. However, as the next example demonstrates, the same methods can be applied to nonlinear models, such as the discrete choice models. 15 16.10

17-72

We have examined two approaches to estimation of a probit model with random effects. GMM estimation is another possibility. Avery, Hansen, and Hotz (1983), Bertschek and Lechner (1998), and Inkmann (2000) examine this approach; the latter two offer some comparison with the quadrature and simulation-based estimators considered here. ~~(Our application in Example 23.16 will use the Bertschek and Lechner data.)~~

~~The preceding opens another possibility.~~ The random effects model can be cast as a model with a random constant term;

$$y_{it}^* = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, \; t = 1, \ldots, T_i,$$
$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,}$$

where $\alpha_i = \alpha + \sigma_u u_i$. This is simply a reinterpretation of the model we just analyzed. We might, however, now extend this formulation to the full parameter vector. The resulting structure is

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}, \quad i = 1, \ldots, n, \; t = 1, \ldots, T_i,$$
$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,}$$

AU: This was a KT on msp 17-51. Here also?

where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_i$ where $\boldsymbol{\Gamma}$ is a nonnegative definite diagonal matrix—some of its diagonal elements could be zero for nonrandom parameters. The method of estimation is maximum simulated likelihood. ~~is essentially the same as before~~. The simulated log-likelihood is now

$$\ln L_{Simulated} = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{t=1}^{T_i} F[q_{it}(\mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_{ir}))] \right] \right\}.$$

The simulation now involves R draws from the multivariate distribution of $\mathbf{u}$. Because the draws are uncorrelated—$\boldsymbol{\Gamma}$ is diagonal—this is essentially the same estimation problem as the random effects model considered previously. This model is estimated in Example ~~23.11~~ 17.16. Example ~~23.11~~ 17.16 presents a similar model that assumes that the distribution of $\boldsymbol{\beta}_i$ is discrete rather than continuous.

23.5.2 Fixed Effects Models

The fixed effects model is

$$y_{it}^* = \alpha_i d_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, \; t = 1, \ldots, T_i,$$
$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,}$$

where d_{it} is a dummy variable that takes the value one for individual i and zero otherwise. For convenience, we have redefined $\mathbf{x}_{it}$ to be the nonconstant variables in the model. The parameters to be estimated are the K elements of $\boldsymbol{\beta}$ and the n individual constant terms. Before we consider the several virtues and shortcomings of this model, we consider the practical aspects of estimation of what are possibly a huge number of parameters, $(n + K) - n$ is not limited here, and could be in the thousands in a typical application. The log-likelihood function for the fixed effects model is

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \ln P(y_{it} \,|\, \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}),$$

where $P(.)$ is the probability of the observed outcome, for example, $\Phi[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$ for the probit model or $\Lambda[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$ for the logit model. What follows can be extended to any index function model, but for the present, we'll confine our attention

17-73

TABLE ~~23.9~~ 17.13 Estimated Heterogeneous Parameter Models

	Pooled	*Random Parameters*		*Latent Class*		
Variable	*Estimate: β*	*Estimate: β*	*Estimate: σ*	*Estimate: β*	*Estimate: β*	*Estimate: β*
Constant	0.25111 (0.091135)	−0.034964 (0.075533)	0.81651 (0.016542)	0.96605 (0.43757)	−0.18579 (0.23907)	−1.52595 (0.43498)
Age	0.020709 (0.0012852)	0.026306 (0.0011038)	0.025330 (0.0004226)	0.049058 (0.0069455)	0.032248 (0.0031462)	0.019981 (0.0062550)
Income	−0.18592 (0.075064)	−0.0043649 (0.062445)	0.10737 (0.038276)	−0.27917 (0.37149)	−0.068633 (0.16748)	0.45487 (0.31153)
Kids	−0.22947 (0.029537)	−0.17461 (0.024522)	0.55520 (0.023866)	−0.28385 (0.14279)	−0.28336 (0.066404)	−0.11708 (0.12363)
Education	−0.045588 (0.0056465)	−0.040510 (0.0047520)	0.037915 (0.0013416)	−0.025301 (0.027768)	−0.057335 (0.012465)	−0.09385 (0.027965)
Married	0.085293 (0.033286)	0.014618 (0.027417)	0.070696 (0.017362)	−0.10875 (0.17228)	0.025331 (0.075929)	0.23571 (0.14369)
Class Prob.	1.00000 (0.00000)	1.00000 (0.00000)		0.34833 (0.038495)	0.46181 (0.028062)	0.18986 (0.022335)
ln *L*	−17673.10	−16271.72			−16265.59	

Example ~~23.11~~ 17.16 Parameter Heterogeneity in a Binary Choice Model

We have extended the logit model for doctor visits from Example ~~23.10~~ 17.11 and 17.15 to allow the parameters to vary randomly across individuals. The random parameters logit model is

$$\text{Prob}(Doctor_{it} = 1) = \Lambda(\beta_{1i} + \beta_{2i}\,Age_{it} + \beta_{3i}\,Income_{it} + \beta_{4i}\,Kids_{it} + \beta_{5i}\,Educ_{it} + \beta_{6i}\,Married_{it}),$$

where the two models for the parameter variation we have employed are:

$$\text{Continuous:}\quad \beta_{ki} = \beta_k + \sigma_k u_{ki},\ u_{ki} \sim N[0, 1],\ k = 1, \ldots, 6,\ \text{Cov}[u_{ki}, u_{mi}] = 0,$$

$$\begin{aligned}\text{Discrete:}\quad \beta_{ki} &= \beta_k^1 \text{ with probability } \pi_1\\ &\ \beta_k^2 \text{ with probability } \pi_2\\ &\ \beta_k^3 \text{ with probability } \pi_3.\end{aligned}$$

TB 17.13 17.~~14~~3

We have chosen a three-class latent class model for the illustration. 17.15 In an application, one might undertake a systematic search, such as in Example ~~23.10,~~ to find a preferred specification. Table ~~23.9~~ presents the fixed parameter (pooled) logit model and the two random parameters versions. (There are infinite variations on these specifications that one might explore—See Chapter ~~17~~ for discussion—we have shown only the simplest to illustrate the models.[28] ~~A more elaborate specification appears in Section 23.11.7.)~~

15

FN 36 36

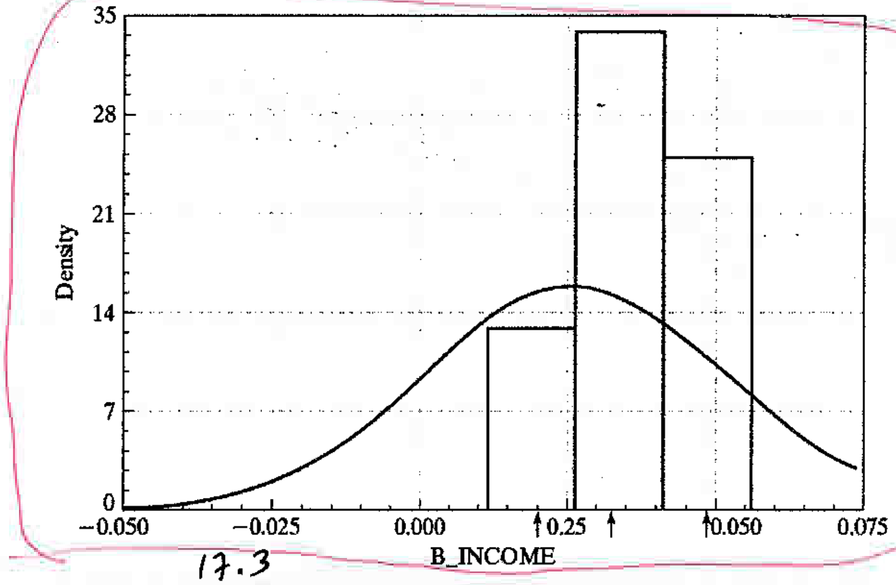
FIG 17.3 17.3

Figure ~~23.3~~ shows the implied distribution for the coefficient on age. For the continuous distribution, we have simply plotted the normal density. For the discrete distribution, we first obtained the mean (0.0358) and standard deviation (0.0107). Notice that the distribution is tighter than the estimated continuous normal (mean, 0.026, standard deviation, 0.0253). To suggest the variation of the parameter (purely for purpose of the display, because the distribution is discrete), we placed the mass of the center interval, 0.462, between the midpoints of the intervals between the center mass point and the two extremes. With a width of 0.0145 the density is 0.461 / 0.0145 = 31.8. We used the same interval widths for the outer segments. This range of variation covers about five standard deviations of the distribution.

36 [28] We have arrived (once again) at a point where the question of replicability 17.12 arises. Nonreplicability is an ongoing challenge in empirical work in economics. (See, e.g., Example ~~23.9~~.) The problem is particularly acute in analyses that involve simulation such as Monte Carlo studies and random parameter models. In the interest of replicability, we note that the random parameter estimates in Table ~~23.9~~ 17.14 were computed with NLOGIT [Econometric Software (2007)] and are based on 50 Halton draws. We used the first six sequences (prime numbers 2, 3, 5, 7, 11, 13) and discarded the first 10 draws in each sequence.

✓

17-74



17.3

FIGURE 23.3 Distributions of Income Coefficient.

~~23.6 SEMIPARAMETRIC ANALYSIS~~

~~In his survey of qualitative response models, Amemiya (1981) reports the following widely cited approximations for the linear probability (LP) model: Over the range of probabilities of 30 to 70 percent,~~

$$\hat{\beta}_{\text{LP}} \approx 0.4\beta_{\text{probit}} \text{ for the slopes,}$$
$$\hat{\beta}_{\text{LP}} \approx 0.25\beta_{\text{logit}} \text{ for the slopes.}$$

~~Aside from confirming our intuition that least squares approximates the nonlinear model and providing a quick comparison for the three models involved, the practical usefulness of the formula is somewhat limited. Still, it is a striking result.[30] A series of studies has focused on reasons why the least squares estimates should be proportional to the probit and logit estimates. A related question concerns the problems associated with assuming that a probit model applies when, in fact, a logit model is appropriate or vice versa.[31] The approximation would seem to suggest that with this type of misspecification, we would once again obtain a scaled version of the correct coefficient vector. (Amemiya also reports the widely observed relationship $\hat{\beta}_{\text{logit}} = 1.6\hat{\beta}_{\text{probit}}$, which follows from the results for the linear probability model.)~~

~~[30] This result does not imply that it is useful to report 2.5 times the linear probability estimates with the probit estimates for comparability. The linear probability estimates are already in the form of marginal effects, whereas the probit coefficients must be scaled *downward*. If the sample proportion happens to be close to 0.5, then the right scale factor will be roughly $\phi[\Phi^{-1}(0.5)] = 0.3989$. But the density falls rapidly as P moves away from 0.5.~~

~~[31] See Ruud (1986) and Gourieroux et al. (1987).~~

17-75

17.4.9 Nonresponse, Attrition and Inverse Probability Weighting

Missing observations is a common problem in the analysis of panel data. Nicoletti and Peracchi (2005) suggest several reasons that, for example, panels become unbalanced:

- Demographic events such as death;
- Movement out of the scope of the survey, such as institutionalization or emigration;
- Refusal to respond at subsequent waves;
- Absence of the person at the address;
- Other types of non-contact.

The GSOEP that we (from Riphahn, Wambach and Million (2003)) have used in many examples in this text is one such data set. Jones, Koolman and Rice (2006) (JKR) list several other applications, including the British Household Panel Survey (BHPS), the European Community Household Panel (ECHP) and the Panel Study of Income Dynamics (PSID).

If observations are missing completely at random (MCAR), then the problem of nonresponse can be ignored, though for estimation of dynamic models, either the analysis will have to be restricted to observations with uninterrupted sequences of observations, or some very strong assumptions and interpolation methods will have to be employed to fill the gaps. (See Section 4.7.4 for discussion of the terminology and issues in handling missing data.) The problem for estimation arises when observations are missing for reasons that are related to the outcome variable of interest. "**Nonresponse bias**" and a related problem, "**attrition bias**" (individuals leave permanently during the study) result when conventional estimators, such as least squares or the probit maximum likelihood estimator being used here, are applied to samples in which observations are present or absent from the sample for reasons related to the outcome variable. It is a form of "**sample selection bias**," that we will examine further in Chapter 19. KT

Verbeek and Nijman (1992) have suggested a test for endogeneity of the sample response pattern. (We will adopt JKR's notation and terminology for this.) Let h denote the outcome of interest and $\mathbf{x}$ denote the relevant set of covariates. Let R denote the pattern of response. If nonresponse is (completely) random, then $E[h|\mathbf{x},R] = E[h|\mathbf{x}]$. This suggests a variable addition test (neglecting other panel data effects); a pooled model that contains R in addition to $\mathbf{x}$ can provide the means for a simple test of endogeneity. JKR (and Verbeek and Nijman) suggest using the number of waves at which the individual is present as the measure of R. Thus, adding R to the pooled model, we can use a simple t test for the hypothesis.

AU: This term not in chap. list

Devising an estimator given that (non)response is nonignorable requires a more detailed understanding of the process generating the response pattern. The crucial issue is whether the sample selection is based "on unobservables" or "on observables." **Selection on unobservables** results when, after conditioning on the relevant variables, $\mathbf{x}$ and other information, $\mathbf{z}$, the sampling mechanism is still nonrandom with respect to the disturbances in the models. Selection on unobservables is at the heart of the sample selectivity methodology pioneered by Heckman (1979) that we will study in Chapter 19. (Some applications of the role of unobservables in biased estimation are discussed in Chapter 8, where we examine sources of endogeneity in regression models.) If selection is on observables, and then conditioned on an appropriate specification involving the observable information, $(\mathbf{x},\mathbf{z})$, a consistent estimator of the model parameters will be available by "purging" the estimator of the endogeneity of the sampling mechanism. KT

KT

JKR adopt an **inverse probability weighted** (IPW) estimator devised by Robins, Rotnitsky, and Zhao (1995), Fitzgerald, Gottshalk, and Moffitt (1998), Moffitt, Fitzgerald, and Gottshalk (1999), and Wooldridge (2002). The estimator is based on the general MCAR assumption that $P(R = 1|h,\mathbf{x},\mathbf{z}) = P(R=1|\mathbf{x},\mathbf{z})$. That is, the observable covariates convey all the information that determines the response pattern – the probability of nonresponse does not vary

17-76

systematically with the outcome variable once the exogenous information is accounted for. Implementing this idea in an estimator would require that **x** and **z** be observable when $R = 0$, that is, the exogenous data be available for the nonresponders. This will typically not be the case; in an unbalanced panel, the entire observation is missing. Wooldridge (2002) proposed somewhat stronger assumption that makes estimation feasible; $P(R = 1|h,\mathbf{x},\mathbf{z}) = P(R = 1|\mathbf{z})$ where **z** is a set of covariates available at wave 1 (entry to the study). To compute Wooldridge's IPW estimator, we will begin with the sample of all individuals who are present at wave 1 of the study. (In our example 17.17, based on the GSOEP data, not all individuals are present at the first wave.) At wave 1, $(\mathbf{x}_{i1},\mathbf{z}_{i1})$ are observed for all individuals to be studied; $\mathbf{z}_{i1}$ contains information on observables that are not included in the outcome equation and that predict the response pattern at subsequent waves, including the response variable at the first wave. At wave 1, then, $P(R_{i1}=1|\mathbf{x}_{i1},\mathbf{z}_{i1}) = 1$. Wooldridge suggests using a probit model for $P(R_{it} = 1|\mathbf{x}_{i1},\mathbf{z}_{i1})$, $t = 2,\ldots,T$ for the remaining waves to obtain predicted probabilities of response, $\hat{p}_{it}$. The IPW estimator then maximizes the weighted log likelihood

$$\ln L_{IPW} = \sum\nolimits_{i=1}^{n} \sum\nolimits_{t=1}^{T} \frac{R_{it}}{\hat{p}_{it}} \ln L_{it}\,.$$

Inference based on the weighted log likelihood function can proceed as in Section 17.3. A remaining detail concerns whether the use of the predicted probabilities in the weighted log likelihood function makes it necessary to correct the standard errors for two step estimation. The case here is not an application of the two step estimators we considered in Section 14.7, since the first step is not used to produce an estimated parameter vector in the second. Wooldridge (2002) shows that the standard errors computed without the adjustment are "conservative" in that they are larger than they would be with the adjustment.

Example 17.17 Nonresponse in the GSOEP Sample

FIG 17.4

Of the 7,293 individuals in the GSOEP data that we have used in several earlier examples, 3,874 were present at wave 1 (1984) of the sample. The pattern of the number of waves present by these 3,874 are shown in Figure 17.4. The waves are 1984-1988, 1991 and 1994. A dynamic model would be based on the 1,600 of those present at wave 1 who were also present for the next four waves. There is a substantial amount of nonresponse in these data. Not all individuals exit the sample with the first nonresponse, however, so the resulting panel remains unbalanced. The impression suggested by Figure 17.4 could be a bit misleading – the nonresponse pattern is quite different from simple attrition. For example, of the 3,874 individuals who responded at wave 1, 364 did not respond at wave 2 but returned to the sample at wave 3.

TB 17.14

To employ the Verbeek and Nijman test, we used the entire sample of 27,326 household years of data. The pooled probit model for DocVis > 0 produced the results at the left in table 17.14. A *t* (Wald) test of the hypothesis that the coefficient on number of waves present is zero is strongly rejected, so we proceed to the inverse probability weighted estimator. For computing the inverse probability weights, we used the following specification:

x_{i1} = *constant, age, income, educ, kids, married*
z_{i1} = *female, handicapped dummy, percentage handicapped,*
university, working, blue collar, white collar, public servant, y_{i1}
y_{i1} = *Doctor Visits* > 0 in period 1.

This first year data vector is used as the observed explanatory variables in probit models for waves 2-7 for the 3,874 individuals who were present at wave 1. There are 3,874 observations for each of these probit models, since all were observed at wave 1. Fitted probabilities for R_{it} are computed for waves 2-7, while $R_{i1} = 1$. The sample means of these

17-77

? s who

probabilities which equal the proportion of the 3,874 that responded at each wave are 1.000, 0.730, 0.672, 0.626, 0.682, 0.568, and 0.386, respectively. Table 17.14 presents the estimated models for several specifications In each case, it appears that the weighting brings some moderate changes in the parameters and, uniformly, reductions in the standard errors.

4

a

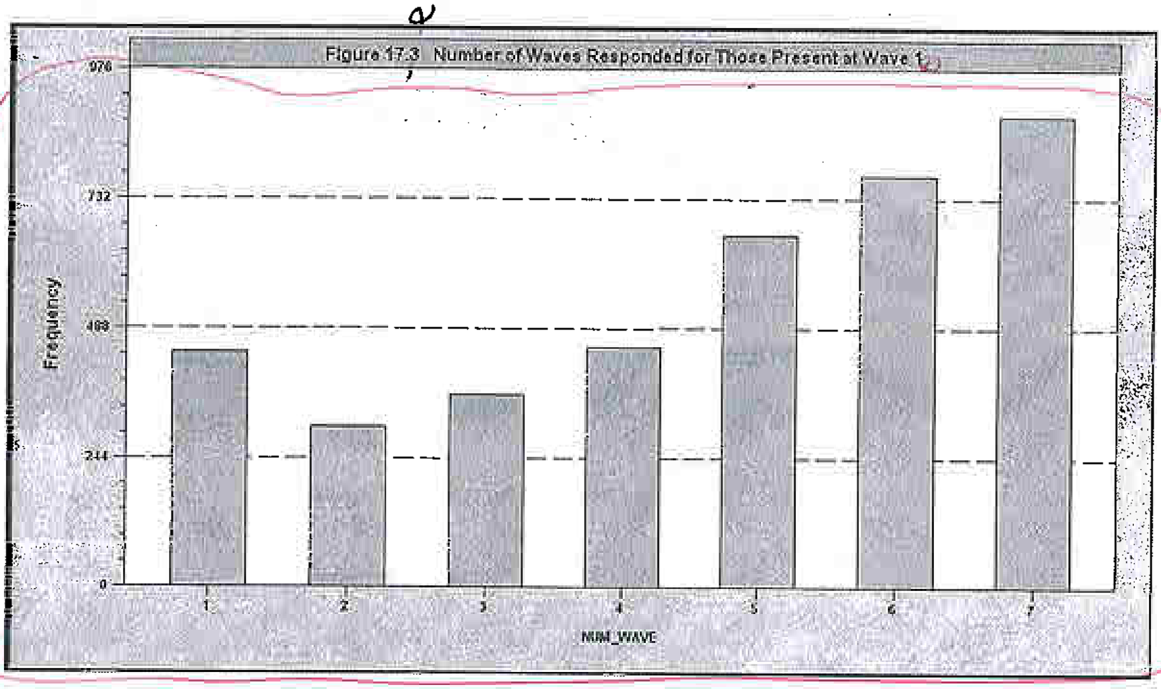


TABLE 17.14 Inverse Probability Weighted Estimators

Variable	Endog. Test	Pooled Model		Random Effects - M Mundlak		Fixed Effects	
		Unwtd.	IPW	Unwtd.	IPW	Unwtd.	IPW
Constant	0.26411 (0.05893)	0.03369 (0.07684)	-0.02373 (0.06385)	0.09838 (0.16081)	0.13237 (0.17019)		
Age	0.01369 (0.00080)	0.01667 (0.00107)	0.01831 (0.00088)	0.05141 (0.00422)	0.05656 (0.00388)	0.06210 (0.00506)	0.06841 (0.00465)
Income	-0.12446 (0.04636)	-0.17097 (0.05981)	-0.22263 (0.04801)	0.05794 (0.11256)	0.01699 (0.10580)	0.07880 (0.12891)	0.03603 (0.12193)
Education	-0.02925 (0.00351)	-0.03614 (0.00449)	-0.03513 (0.00365)	-0.06456 (0.06104)	-0.07058 (0.05792)	-0.07752 (0.06582)	-0.08574 (0.06149)
Kids	-0.13130 (0.01828)	-0.13077 (0.02303)	-0.13277 (0.01950)	-0.04961 (0.04500)	-0.03427 (0.04356)	-0.05776 (0.05296)	-0.03546 (0.05166)
Married	0.06759 (0.02060)	0.06237 (0.02616)	0.07015 (0.02097)	-0.06582 (0.06596)	-0.09235 (0.06330)	-0.07939 (0.08146)	-0.11283 (0.07838)
Mean Age				-0.03056 (0.00479)	-0.03401 (0.00455)		
Mean Income				-0.66388 (0.18646)	-0.78077 (0.18866)		
Mean Education				0.02656 (0.06160)	0.02899 (0.05848)		
Mean Kids				-0.17524 (0.07266)	-0.20615 (0.07464)		
Mean Married				0.22346 (0.08719)	0.25763 (0.08433)		
Number of Waves	-0.02977 (0.00450)						
ρ				0.46538	0.48616		

17.5 BIVARIATE AND MULTIVARIATE PROBIT MODELS

In Chapter 10, we analyzed a number of different multiple-equation extensions of the classical and generalized regression model. A natural extension of the probit model would be to allow more than one equation, with correlated disturbances, in the same spirit as the seemingly unrelated regressions model. The general specification for a two-equation model would be

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad y_1 = 1 \text{ if } y_1^* > 0, \text{ 0 otherwise}, \\ y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y_2 = 1 \text{ if } y_2^* > 0, \text{ 0 otherwise}, \\ &\begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix}\Bigg| \mathbf{x}_1,\mathbf{x}_2 \sim N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right]. \end{aligned} \tag{17-49}$$

48

This bivariate probit model is interesting in its own right for modeling the joint determination of two variables, such as doctor and hospital visits in the next example. It also provides the framework for modeling in two common applications. In many cases, a treatment effect, or endogenous influence takes place in a binary choice context. The bivariate probit model provides a specification for analyzing a case in which a probit model contains an endogenous binary variable in one of the equations. In example 17.21, we will extend (17-49) to c48

$$\begin{aligned} W^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad W = 1 \text{ if } W^* > 0, \text{ 0 otherwise}, \\ y^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \gamma W + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, \text{ 0 otherwise}, \\ &\begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix}\Bigg| \mathbf{x}_1,\mathbf{x}_2 \sim N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right]. \end{aligned} \tag{17-50}$$

49

This model extends the case in Section 17.3.5, where *W**, rather than *W*, appears on the right hand side of the second equation. In the example, *W* denotes whether a liberal arts college supports a women's studies program on the campus while y is a binary indicator of whether the economics department provides a gender economics course. A second common application, in which the first equation is an endogenous sampling rule, is another variant of the bivariate probit model;

$$\begin{aligned} S^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad S = 1 \text{ if } S^* > 0, \text{ 0 otherwise}, \\ y^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, \text{ 0 otherwise}, \\ &\begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix}\Bigg| \mathbf{x}_1,\mathbf{x}_2 \sim N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right], \\ &(y\ ,\mathbf{x}_2) \text{ observed only when } S = 1. \end{aligned} \tag{17-51}$$

50

In Example 17.22, we will study an application in which *S* is the result of a credit card application (or any sort of loan application) while y_2 is a binary indicator for whether the individual defaults on the credit account (loan). This is a form of endogenous sampling (in this instance, sampling on unobservables) that has some commonality with the attrition problem that we encountered in Section 17.4.9.

48

At the end of this section, we will extend (17-49) to more than two equations. This will allow direct treatment of multiple binary outcomes. It will also allow a more general panel data model for *T* periods than is provided by the random effects specification.

17-79

approach to estimating the structural parameters. In an application somewhat similar to Example 23.13, they apply the technique to a labor force participation model for British men in which a variable of interest is a dummy variable for education greater than 16 years, the endogenous variable in the participation equation, also of interest, is earned income of the spouse, and an instrumental variable is a welfare benefit entitlement. Their findings are rather more substantial than ours; they find that when the endogeneity of other family income is accommodated in the equation, the education coefficient increases by 40 percent and remains significant, but the coefficient on other income increases by more than tenfold.

The case in which the endogenous variable in the main equation is, itself, a binary variable occupies a large segment of the recent literature. Consider the model

$$
\begin{aligned}
y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + \varepsilon_i,\\
y_i &= 1(y_i^* > 0),\\
E[\varepsilon_i \mid T_i] &\neq 0,
\end{aligned}
$$

where T_i is a binary variable indicating some kind of program participation (e.g., graduating from high school or college, receiving some kind of job training, etc.). The model in this form (and several similar ones) is a "treatment effects" model. The main object of estimation is γ (at least superficially). In these settings, the observed outcome may be y_i^* (e.g., income or hours) or y_i (e.g., labor force participation). The preceding analysis has suggested that problems of endogeneity will intervene in either case. The subject of treatment effects models is surveyed in many studies, including Angrist (2001). We will examine this model in some detail in Chapter 24.

17.5 ~~23.8~~ BIVARIATE PROBIT MODELS AND MULTIVARIATE

In Chapter 10, we analyzed a number of different multiple-equation extensions of the classical and generalized regression model. A natural extension of the probit model would be to allow more than one equation, with correlated disturbances, in the same spirit as the seemingly unrelated regressions model. The general specification for a two-equation model would be

$$
\begin{aligned}
y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad y_1 = 1 \quad \text{if } y_1^* > 0, 0 \text{ otherwise},\\
y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y_2 = 1 \quad \text{if } y_2^* > 0, 0 \text{ otherwise},\\
E[\varepsilon_1 \mid \mathbf{x}_1, \mathbf{x}_2] &= E[\varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = 0,\\
\mathrm{Var}[\varepsilon_1 \mid \mathbf{x}_1, \mathbf{x}_2] &= \mathrm{Var}[\varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = 1,\\
\mathrm{Cov}[\varepsilon_1, \varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] &= \rho.
\end{aligned}
$$

17-49 ~~(23-44)~~

17.5.1 ~~23.8.1~~ MAXIMUM LIKELIHOOD ESTIMATION

The bivariate normal cdf is

$$
\mathrm{Prob}(X_1 < x_1, X_2 < x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} \phi_2(z_1, z_2, \rho)\, dz_1 dz_2,
$$

17- 80

FN 37

which we denote $\Phi_2(x_1, x_2, \rho)$. The density is~~[39]~~ 37

$$\phi_2(x_1, x_2, \rho) = \frac{e^{-(1/2)(x_1^2+x_2^2-2\rho x_1 x_2)/(1-\rho^2)}}{2\pi(1-\rho^2)^{1/2}}.$$

To construct the log-likelihood, let $q_{i1} = 2y_{i1} - 1$ and $q_{i2} = 2y_{i2} - 1$. Thus, $q_{ij} = 1$ if $y_{ij} = 1$ and -1 if $y_{ij} = 0$ for $j = 1$ and 2. Now let

$$z_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}_j \quad \text{and} \quad w_{ij} = q_{ij}z_{ij}, \quad j = 1, 2,$$

and

$$\rho_{i*} = q_{i1}q_{i2}\rho.$$

Note the notational convention. The subscript 2 is used to indicate the bivariate normal distribution in the density ϕ_2 and cdf Φ_2. In all other cases, the subscript 2 indicates the variables in the second equation. As before, $\phi(.)$ and $\Phi(.)$ without subscripts denote the univariate standard normal density and cdf.

The probabilities that enter the likelihood function are

$$\text{Prob}(Y_1 = y_{i1}, Y_2 = y_{i2} \mid \mathbf{x}_1, \mathbf{x}_2) = \Phi_2(w_{i1}, w_{i2}, \rho_{i*}),$$

FN 38

which accounts for all the necessary sign changes needed to compute probabilities for y's equal to zero and one. Thus,~~[40]~~ 38

$$\ln L = \sum_{i=1}^{n} \ln \Phi_2(w_{i1}, w_{i2}, \rho_{i*}).$$

The derivatives of the log-likelihood then reduce to

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{n}\left(\frac{q_{ij}g_{ij}}{\Phi_2}\right)\mathbf{x}_{ij}, \quad j = 1, 2,$$

17-~~49~~ 51 (~~23-45~~) 52

$$\frac{\partial \ln L}{\partial \rho} = \sum_{i=1}^{n}\frac{q_{i1}q_{i2}\phi_2}{\Phi_2},$$

53 17-~~50~~ 52 (~~23-46~~)

where

$$g_{i1} = \phi(w_{i1})\Phi\left[\frac{w_{i2} - \rho_{i*}w_{i1}}{\sqrt{1-\rho_{i*}^2}}\right]$$

and the subscripts 1 and 2 in g_{i1} are reversed to obtain g_{i2}. Before considering the Hessian, it is useful to note what becomes of the preceding if $\rho = 0$. For $\partial \ln L/\partial \boldsymbol{\beta}_1$, if $\rho = \rho_{i*} = 0$, then g_{i1} reduces to $\phi(w_{i1})\Phi(w_{i2})$, ϕ_2 is $\phi(w_{i1})\phi(w_{i2})$, and Φ_2 is $\Phi(w_{i1})\Phi(w_{i2})$. Inserting these results in ~~(23-45)~~ 17-~~50~~ 51 with q_{i1} and q_{i2} produces ~~(23-21)~~ 17-21. Because both functions in $\partial \ln L/\partial \rho$ factor into the product of the univariate functions, $\partial \ln L/\partial \rho$ reduces to $\sum_{i=1}^{n} \lambda_{i1}\lambda_{i2}$, where λ_{ij}, $j = 1, 2$, is defined in (~~23-21~~ 17-~~21~~ 20). (This result will reappear in the LM statistic shown later.)

The maximum likelihood estimates are obtained by simultaneously setting the three derivatives to zero. The second derivatives are relatively straightforward but tedious.

37 ~~[39]~~See Section B.9.

38 ~~[40]~~To avoid further ambiguity, and for convenience, the observation subscript will be omitted from $\Phi_2 = \Phi_2(w_{i1}, w_{i2}, \rho_{i*})$ and from $\phi_2 = \phi_2(w_{i1}, w_{i2}, \rho_{i*})$.

17-81

Some,simplifications are useful. Let

$$\delta_i = \frac{1}{\sqrt{1-\rho_{i*}^2}},$$
$$v_{i1} = \delta_i(w_{i2} - \rho_{i*} w_{i1}), \quad \text{so } g_{i1} = \phi(w_{i1})\Phi(v_{i1}),$$
$$v_{i2} = \delta_i(w_{i1} - \rho_{i*} w_{i2}), \quad \text{so } g_{i2} = \phi(w_{i2})\Phi(v_{i2}).$$

By multiplying it out, you can show that

$$\delta_i\phi(w_{i1})\phi(v_{i1}) = \delta_i\phi(w_{i2})\phi(v_{i2}) = \phi_2.$$

Then

$$\frac{\partial^2 \log L}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1'} = \sum_{i=1}^n \mathbf{x}_{i1}\mathbf{x}_{i1}'\left[\frac{-w_{i1}g_{i1}}{\Phi_2} - \frac{\rho_{i*}\phi_2}{\Phi_2} - \frac{g_{i1}^2}{\Phi_2^2}\right],$$
$$\frac{\partial^2 \log L}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_2'} = \sum_{i=1}^n q_{i1}q_{i2}\mathbf{x}_{i1}\mathbf{x}_{i2}'\left[\frac{\phi_2}{\Phi_2} - \frac{g_{i1}g_{i2}}{\Phi_2^2}\right],$$
$$\frac{\partial^2 \log L}{\partial\boldsymbol{\beta}_1\partial\rho} = \sum_{i=1}^n q_{i2}\mathbf{x}_{i1}\frac{\phi_2}{\Phi_2}\left[\rho_{i*}\delta_i v_{i1} - w_{i1} - \frac{g_{i1}}{\Phi_2}\right],$$
$$\frac{\partial^2 \log L}{\partial\rho^2} = \sum_{i=i}^n \frac{\phi_2}{\Phi_2}\left[\delta_i^2\rho_{i*}(1 - \mathbf{w}_i'\mathbf{R}_i^{-1}\mathbf{w}_i) + \delta_i^2 w_{i1}w_{i2} - \frac{\phi_2}{\Phi_2}\right],$$

53

17-~~52~~ 54
(~~23=47~~)

where $\mathbf{w}_i'\mathbf{R}_i^{-1}\mathbf{w}_i = \delta_i^2(w_{i1}^2 + w_{i2}^2 - 2\rho_{i*}w_{i1}w_{i2})$. (For $\boldsymbol{\beta}_2$, change the subscripts in $\partial^2 \ln L/\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1'$ and $\partial^2 \ln L/\partial\boldsymbol{\beta}_1\partial\rho$ accordingly.) The complexity of the second derivatives for this model makes it an excellent candidate for the Berndt et al. estimator of the variance matrix of the maximum likelihood estimator.

Example ~~23.14~~ 17.18 Tetrachoric Correlation

17.4 and several others,

Returning once again to the health care application of Examples 1~~1.11, 16.16, 23.4, and 23.8~~, we now consider a second binary variable,

$$\textit{Hospital}_{it} = 1 \text{ if } \textit{HospVis}_{it} > 0 \text{ and } 0 \text{ otherwise.}$$

Our previous analyses have focused on

$$\textit{Doctor}_{it} = 1 \text{ if } \textit{DocVis}_{it} > 0 \text{ and } 0 \text{ otherwise.}$$

A simple bivariate frequency count for these two variables is

	Hospital		
Doctor	***0***	***1***	***Total***
0	9,715	420	10,135
1	15,216	1,975	17,191
Total	24,931	2,395	27,326

Looking at the very large value in the lower-left cell, one might surmise that these two binary variables (and the underlying phenomena that they represent) are negatively correlated. The usual Pearson, product moment correlation would be inappropriate as a measure of this correlation since it is used for continuous variables. Consider, instead, a bivariate probit "model,"

$$H_{it}^* = \mu_1 + \varepsilon_{1,it}, \quad \textit{Hospital}_{it} = 1(H_{it}^* > 0),$$
$$D_{it}^* = \mu_2 + \varepsilon_{2,it}, \quad \textit{Doctor}_{it} = 1(D_{it}^* > 0),$$

17-~~49~~ 48

where $(\varepsilon_1, \varepsilon_2)$ have a bivariate normal distribution with means (0, 0), variances (1, 1) and correlation ρ. This is the model in (~~23-44~~) without independent variables. In this representation, KT the **tetrachoric correlation,** which is a correlation measure for a pair of binary variables, is precisely the ρ in this model—it is the correlation that would be measured between the underlying continuous variables if they could be observed. This suggests an interpretation of the correlation coefficient in a bivariate probit model—as the conditional tetrachoric correlation. It also suggests a method of easily estimating the tetrachoric correlation coefficient using a program that is built into nearly all commercial software packages.

Applied to the hospital/doctor data defined earlier, we obtained an estimate of ρ of 0.31106, with an estimated asymptotic standard error of 0.01357. Apparently, our earlier intuition was incorrect.

17.5.2 ~~23.8.2~~ TESTING FOR ZERO CORRELATION

The Lagrange multiplier statistic is a convenient device for testing for the absence of correlation in this model. Under the null hypothesis that ρ equals zero, the model consists of independent probit equations, which can be estimated separately. Moreover, in the multivariate model, all the bivariate (or multivariate) densities and probabilities factor into the products of the marginals if the correlations are zero, which makes construction of the test statistic a simple matter of manipulating the results of the independent probits. The Lagrange multiplier statistic for testing $H_0\colon \rho = 0$ in a bivariate probit model is[~~44~~][39] FN 39

$$\mathrm{LM} = \frac{\left[\sum_{i=1}^{n} q_{i1}q_{i2}\dfrac{\phi(w_{i1})\phi(w_{i2})}{\Phi(w_{i1})\Phi(w_{i2})}\right]^2}{\sum_{i=1}^{n}\dfrac{[\phi(w_{i1})\phi(w_{i2})]^2}{\Phi(w_{i1})\Phi(-w_{i1})\Phi(w_{i2})\Phi(-w_{i2})}}.$$

As usual, the advantage of the LM statistic is that it obviates computing the bivariate probit model. But the full unrestricted model is now fairly common in commercial software, so that advantage is minor. The likelihood ratio or Wald test can often be used with equal ease. To carry out the likelihood ratio test, we note first that if ρ equals zero, then the bivariate probit model becomes two independent univariate probits models. The log-likelihood in that case would simply be the sum of the two separate log-likelihoods. The test statistic would be

$$\lambda_{\mathrm{LR}} = 2[\ln L_{\mathrm{BIVARIATE}} - (\ln L_1 + \ln L_2)].$$

This would converge to a chi-squared variable with one degree of freedom. The Wald test is carried out by referring

$$\lambda_{\mathrm{WALD}} = \left[\hat{\rho}_{MLE}\big/\sqrt{\text{Est. Asy. Var}[\hat{\rho}_{MLE}]}\right]^2$$

to the chi-squared distribution with one degree of freedom. For 95 percent significance, the critical value is 3.84 (or one can refer the positive square root to the standard normal critical value of 1.96). Example ~~23.14~~ 17-19 demonstrates.

[39] [~~41~~] This is derived in Kiefer (1982).

17.5.3 ~~23.8.3~~ PARTIAL ~~MARGINAL~~ EFFECTS

FN 40

40 49 17-50 There are several "marginal effects" one might want to evaluate in a bivariate probit model.~~42~~ A natural first step would be the derivatives of Prob$[y_1 = 1, y_2 = 1 \mid \mathbf{x}_1, \mathbf{x}_2]$. These can be deduced from (~~23-45~~) by multiplying by Φ_2, removing the sign carrier, q_{ij} and differentiating with respect to $\mathbf{x}_j$ rather than $\boldsymbol{\beta}_j$. The result is

$$\frac{\partial \Phi_2(\mathbf{x}_1'\boldsymbol{\beta}_1, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho)}{\partial \mathbf{x}_1} = \phi(\mathbf{x}_1'\boldsymbol{\beta}_1)\Phi\left(\frac{\mathbf{x}_2'\boldsymbol{\beta}_2 - \rho\mathbf{x}_1'\boldsymbol{\beta}_1}{\sqrt{1-\rho^2}}\right)\boldsymbol{\beta}_1.$$

Note, however, the bivariate probability, albeit possibly of interest in its own right, is not a conditional mean function. As such, the preceding does not correspond to a regression coefficient or a slope of a conditional expectation.

For convenience in evaluating the conditional mean and its partial effects, we will define a vector $\mathbf{x} = \mathbf{x}_1 \cup \mathbf{x}_2$ and let $\mathbf{x}_1'\boldsymbol{\beta}_1 = \mathbf{x}'\boldsymbol{\gamma}_1$. Thus, $\boldsymbol{\gamma}_1$ contains all the nonzero elements of $\boldsymbol{\beta}_1$ and possibly some zeros in the positions of variables in $\mathbf{x}$ that appear only in the other equation; $\boldsymbol{\gamma}_2$ is defined likewise. The bivariate probability is

$$\text{Prob}[y_1 = 1, y_2 = 1 \mid \mathbf{x}] = \Phi_2[\mathbf{x}'\boldsymbol{\gamma}_1, \mathbf{x}'\boldsymbol{\gamma}_2, \rho].$$

Signs are changed appropriately if the probability of the zero outcome is desired in either case. (See ~~23-44.~~) 17-49 18 The marginal effects of changes in $\mathbf{x}$ on this probability are given by

$$\frac{\partial \Phi_2}{\partial \mathbf{x}} = g_1\boldsymbol{\gamma}_1 + g_2\boldsymbol{\gamma}_2, \qquad \text{17-51 50}$$

where g_1 and g_2 are defined in (~~23-46~~). The familiar univariate cases will arise if $\rho = 0$, and effects specific to one equation or the other will be produced by zeros in the corresponding position in one or the other parameter vector. There are also some conditional mean functions to consider. The unconditional mean functions are given by the univariate probabilities:

$$E[y_j \mid \mathbf{x}] = \Phi(\mathbf{x}'\boldsymbol{\gamma}_j), \quad j = 1, 2,$$

so the analysis of (~~23~~ 17-9) and (~~23~~ 17-10) applies. One pair of conditional mean functions that might be of interest are

$$E[y_1 \mid y_2 = 1, \mathbf{x}] = \text{Prob}[y_1 = 1 \mid y_2 = 1, \mathbf{x}] = \frac{\text{Prob}[y_1 = 1, y_2 = 1 \mid \mathbf{x}]}{\text{Prob}[y_2 = 1 \mid \mathbf{x}]}$$

$$= \frac{\Phi_2(\mathbf{x}'\boldsymbol{\gamma}_1, \mathbf{x}'\boldsymbol{\gamma}_2, \rho)}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}$$

and similarly for $E[y_2 \mid y_1 = 1, \mathbf{x}]$. The marginal effects for this function are given by

$$\frac{\partial E[y_1 \mid y_2 = 1, \mathbf{x}]}{\partial \mathbf{x}} = \left(\frac{1}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}\right)\left[g_1\boldsymbol{\gamma}_1 + \left(g_2 - \Phi_2\frac{\phi(\mathbf{x}'\boldsymbol{\gamma}_2)}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}\right)\boldsymbol{\gamma}_2\right].$$

Finally, one might construct the nonlinear conditional mean function

$$E[y_1 \mid y_2, \mathbf{x}] = \frac{\Phi_2[\mathbf{x}'\boldsymbol{\gamma}_1, (2y_2 - 1)\mathbf{x}'\boldsymbol{\gamma}_2, (2y_2 - 1)\rho]}{\Phi[(2y_2 - 1)\mathbf{x}'\boldsymbol{\gamma}_2]}.$$

The derivatives of this function are the same as those presented earlier, with sign changes in several places if $y_2 = 0$ is the argument.

[~~42~~ 40] See Greene (1996b) and Christofides et al. (1997, 2000).

17-84

~~The derivatives of this function are the same as those presented earlier, with sign changes in several places if $y_2 = 0$ is the argument.~~

17.~~20~~ 19

Example ~~23.15~~ Bivariate Probit Model for Health Care Utilization

We have extended the bivariate probit model of the previous example by specifying a set of independent variables,

$$\mathbf{x}_i = Constant, Female_i, Age_{it}, Income_{it}, Kids_{it}, Education_{it}, Married_{it}.$$

We have specified that the same exogenous variables appear in both equations. (There is no requirement that different variables appear in the equations, nor that a variable be excluded from each equation.) The correct analogy here is to the seemingly unrelated regressions model, not to the linear simultaneous equations model. Unlike the SUR model of Chapter 10, it is not the case here that having the same variables in the two equations implies that the model can be fit equation by equation, one equation at a time. That result only applies to the estimation of sets of linear regression equations.

QU; x-ref to chap. 10 OK?

17.15 Table ~~23.12~~ contains the estimates of the parameters of the univariate and bivariate probit models. The tests of the null hypothesis of zero correlation strongly reject the hypothesis that ρ equals zero. The t statistic for ρ based on the full model is $0.2981 / 0.0139 = 21.446$, which is much larger than the critical value of 1.96. For the likelihood ratio test, we compute

TB 17.15

$$\lambda_{LR} = 2\{-25285.07 - [-17422.72 - (-8073.604)]\} = 422.508.$$

Once again, the hypothesis is rejected. (The Wald statistic is $21.446^2 = 459.957$.) The LM statistic is 383.953. The coefficient estimates agree with expectations. The income coefficient is statistically significant in the doctor equation but not in the hospital equation, suggesting, perhaps, that physican visits are at least to some extent discretionary while hospital visits occur on an emergency basis that would be much less tied to income. The table also contains the decomposition of the partial effects for $E[y_1 \mid y_2 = 1]$. The direct effect is $[g_1 / \Phi(\mathbf{x}'\gamma_2)]\gamma_1$ in the definition given earlier. The mean estimate of $E[y_1 \mid y_2 = 1]$ is 0.821285. In the table in 17.8 Example ~~23.13~~, this would correspond to the raw proportion $P(D = 1, H = 1) / P(H = 1) = (1975 / 27326) / 2395 / 27326) = 0.8246$.

17.15

TABLE ~~23.12~~ Estimated Bivariate Probit Model[a]

	Doctor					Hospital	
	Model Estimates		Partial Effects			Model Estimates	
Variable	Univariate	Bivariate	Direct	Indirect	Total	Univariate	Bivariate
Constant	−0.1243 (0.05815)	−0.1243 (0.05814)				−1.3328 (0.08320)	−1.3385 (0.07957)
Female	0.3559 (0.01602)	0.3551 (0.01604)	0.09650 (0.004957)	−0.00724 (0.001515)	0.08926 (0.005127)	0.1023 (0.02195)	0.1050 (0.02174)
Age	0.01189 (0.0007957)	0.01188 (0.000802)	0.003227 (0.000231)	−0.00032 (0.000073)	0.002909 (0.000238)	0.004605 (0.001082)	0.00461 (0.001058)
Income	−0.1324 (0.04655)	−0.1337 (0.04628)	−0.03632 (0.01260)	−0.003064 (0.004105)	−0.03939 (0.01254)	0.03739 (0.06329)	0.04441 (0.05946)
Kids	−0.1521 (0.01833)	−0.1523 (0.01825)	−0.04140 (0.005053)	0.001047 (0.001773)	−0.04036 (0.005168)	−0.01714 (0.02562)	−0.01517 (0.02570)
Education	−0.01497 (0.003575)	−0.01484 (0.003575)	−0.004033 (0.000977)	0.001512 (0.00035)	−0.002521 (0.0010)	−0.02196 (0.005215)	−0.02191 (0.005110)
Married	0.07352 (0.02064)	0.07351 (0.02063)	0.01998 (0.005626)	0.003303 (0.001917)	0.02328 (0.005735)	−0.04824 (0.02788)	−0.04789 (0.02777)

[a] Estimated correlation coefficient = 0.2981 (0.0139).

17.5.4 A Panel Data Model for Bivariate Binary Response

Extending multiple equation models to accommodate unobserved common effects in panel data settings is straightforward in theory, but complicated in practice. For the bivariate probit case, for example, the natural extension of (17-49) would be -48

$$
\begin{aligned}
y^*_{1,it} &= \mathbf{x}'_{1,it}\boldsymbol{\beta}_1 + \varepsilon_{1,it} + \alpha_{1,i}, \quad y_{1,it} = 1 \text{ if } y^*_{1,it} > 0,\ 0 \text{ otherwise},\\
y^*_{2,it} &= \mathbf{x}'_{2,it}\boldsymbol{\beta}_2 + \varepsilon_{2,it} + \alpha_{2,i}, \quad y_{2,it} = 1 \text{ if } y^*_{2,it} > 0,\ 0 \text{ otherwise},\\
&\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}\Bigg|\mathbf{x}_1,\mathbf{x}_2 \sim N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right].
\end{aligned}
$$

The complication will be in how to treat (α_1,α_2). A fixed effects treatment will require estimation of two full sets of dummy coefficients, will likely encounter the incidental parameters problem in double measure, and will be complicated in practical terms. As in all earlier cases, the fixed effects case also preempts any specification involving time invariant variables. It is also unclear in a fixed effects model, how any correlation between α_1 and α_2 would be handled. It should be noted that strictly from a consistency standpoint, these considerations are moot. The two equations can be estimated separately, only with some loss of efficiency. The analogous situation would be the seemingly unrelated regressions model in chapter 10. A random effects treatment (perhaps accommodated with Mundlak's approach of adding the group means to the equations as in Section 17.4.5) offers greater promise. If $(\alpha_1,\alpha_2) = (u_1,u_2)$ are normally distributed random effects, with

$$
\begin{pmatrix}u_{1,i}\\ u_{2,i}\end{pmatrix}\Bigg|\mathbf{X}_{1,i},\mathbf{X}_{2,i} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right],
$$

then the unconditional log likelihood for the bivariate probit model,

$$
\ln L = \sum_{i=1}^{n} \ln \int_{u_1,u_2} \prod_{t=1}^{T_i} \Phi_2(w_{1,it}\,|\,u_{1,i}, w_{2,it}\,|\,u_{2,i}, \rho^*_{it}) f(u_{1,i},u_{2,i})\,du_{1,i}\,du_{2,i},
$$

can be maximized using simulation or quadrature as we have done in previous applications. A possible variation on this specification would specify that the same common effect enter both equations. In that instance, the integration would only be over a single dimension. In this case, there would only be a single new parameter to estimate, σ^2, the variance of the common random effect while ρ would equal one. A refinement on this form of the model would allow the scaling to be different in the two equations by placing u_i in the first equation and θu_i in the second. This would introduce the additional scaling parameter, but ρ would still equal one. This is the formulation of a common random effect used in Heckman's formulation of the dynamic panel probit model in the Section 17.4.6.

17-86

Example 17.20 Bivariate Random Effects Model for Doctor and Hospital Visits

We will extend the pooled bivariate probit model presented in Example 17.19 by allowing a general random effects formulation, with free correlation between the time varying components ($\varepsilon_1,\varepsilon_2$) and between the time invariant effects, (u_1,u_2). We used simulation to fit the model. Table 17.16 presents the pooled and random effects estimates. The log likelihood functions for the pooled and random effects models are -25285.07 and -23769.67, respectively. Two times the difference is 3030.76. This would be a chi squared with three degrees of freedom (for the three free elements in the covariance matrix of u_1 and u_2). The 95% critical value is 7.81, so the pooling hypothesis would be rejected. The change in the correlation coefficient from .2981 to .1501 suggests that we have decomposed the disturbance in the model into a time varying part and a time invariant part. The latter seems to be the smaller of the two. Although the time invariant elements are more highly correlated, their variances are only $0.2233^2 = 0.0499$ and $0.6338^2 = 0.4017$ compared to 1.0 for both ε_1 and ε_2.

TB 17.16

percent

minus

TABLE 17.16 Estimated Random Effects Bivariate Probit Model

	Doctor		Hospital	
	Pooled	**Random Effects**	**Pooled**	**Random Effects**
Constant	-0.1243 (0.05814)	-0.2976 (0.09650)	-1.3385 (0.07957)	-1.5855 (0.10853)
Female	0.3551 (0.01604)	0.4548 (0.02857)	0.1050 (0.02174)	0.1280 (0.02954)
Age	0.01188 (0.000802)	0.01983 (0.00130)	0.00461 (0.001058)	0.00496 (0.00139)
Income	-0.1337 (0.04628)	-0.01059 (0.06488)	0.04441 (0.05946)	0.13358 (0.07728)
Kids	-0.1523 (0.01825)	-0.1544 (0.02692)	-0.01517 (0.02570)	0.02155 (0.03211)
Education	-0.01484 (0.003575)	-0.02573 (0.00612)	-0.02191 (0.005110)	-0.02444 (0.00675)
Married	0.07351 (0.02063)	0.02876 (0.03167)	-0.04789 (0.02777)	-0.10504 (0.03547)
Corr($\varepsilon_1,\varepsilon_2$)	0.2981	0.1501	0.2981	0.1501
Corr(u_1,u_2)	0.0000	0.5382	0.0000	0.5382
Std. Dev. u	0.0000	0.2233	0.0000	0.6338
Std. Dev. ε	1.0000	1.0000	1.0000	1.0000

17-87

17.5.5 Endogenous Binary Variable a Recursive Bivariate Probit Model

Section 17.3.5 examines a case in which there is an endogenous variable in a binary choice (probit) model. The model is

$$W^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1,$$
$$y^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \gamma W^* + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, \text{ 0 otherwise},$$
$$\begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix} \Big| \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}\right].$$

The application examined there involved a labor force participation model that was conditioned on an endogeous variable, the spouse's hours of work. In many cases, the endogenous variable in the equation is also binary. In the application we will examine ~~below~~ next, the presence of a gender economics course in the economics curriculum at liberal arts colleges is conditioned on whether or not there is a women's studies program on the campus. The model in this case becomes

$$W^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \; W = 1 \text{ if } W^* > 0, \text{ 0 otherwise},$$
$$y^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \gamma W + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, \text{ 0 otherwise},$$
$$\begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix} \Big| \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}\right].$$

This model illustrates a number of interesting aspects of the bivariate probit model. Note that this model is qualitatively different from the bivariate probit model in (17-49); the first dependent variable, W, appears on the right-hand side of the second equation.[41] This model is a **recursive**, simultaneous-equations model. Surprisingly, the endogenous nature of one of the variables on the right-hand side of the second equation can be ignored in formulating the log-likelihood. [The model appears in Maddala (1983, p. 123).] We can establish this fact with the following (admittedly trivial) argument: The term that enters the log-likelihood is $P(y = 1, W = 1) = P(y = 1 \mid W = 1)P(W = 1)$. Given the model as stated, the marginal probability for W is just $\Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)$, whereas the conditional probability is $\Phi_2(\ldots)/\Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)$. The product returns the bivariate normal probability we had earlier. The other three terms in the log-likelihood are derived similarly, which produces (Maddala's results with some sign changes):

= 48

FN 41

KT

$$P(y = 1, W = 1) = \Phi(\mathbf{x}_2'\boldsymbol{\beta}_2 + \gamma, \mathbf{x}_1'\boldsymbol{\beta}_1, \rho),$$
$$P(y = 1, W = 0) = \Phi(\mathbf{x}_2'\boldsymbol{\beta}_2 \qquad , -\mathbf{x}_1'\boldsymbol{\beta}_1, -\rho),$$
$$P(y = 0, W = 1) = \Phi[-(\mathbf{x}_2'\boldsymbol{\beta}_2 + \gamma), \mathbf{x}_1'\boldsymbol{\beta}_1, -\rho),$$
$$P(y = 0, W = 0) = \Phi(-\mathbf{x}_2'\boldsymbol{\beta}_2, -\mathbf{x}_1'\boldsymbol{\beta}_1, \rho).$$

AU: Is something missing in second equation?

[41] Eisenberg and Rowe (2006) is another application of this model. In their study, they analyzed the joint (recursive) effect of W = veteran status on y, smoking behavior. The estimator they used was two-stage least squares and GMM.

17-88

48

These terms are exactly those of (17-49) that we obtain just by carrying W in the second equation with no special attention to its endogenous nature. We can ignore the simultaneity in this model and we cannot in the linear regression model because, in this instance, we are maximizing the log-likelihood, whereas in the linear regression case, we are manipulating certain sample moments that do not converge to the necessary population parameters in the presence of simultaneity.

Example 17.21 Gender Economics Courses at Liberal Arts Colleges

Burnett (1997) proposed the following bivariate probit model for the presence of a gender economics course in the curriculum of a liberal arts college:

$$\text{Prob}[G = 1, W = 1 \mid \mathbf{x}_G, \mathbf{x}_W] = \Phi_2(\mathbf{x}_G'\boldsymbol{\beta}_G + \gamma W, \mathbf{x}_W'\boldsymbol{\beta}_W, \rho).$$

The dependent variables in the model are

G = presence of a gender economics course,
W = presence of a women's studies program on the campus.

The independent variables in the model are

z_1 = constant term;
z_2 = academic reputation of the college, coded 1 (best), 2, . . . to 141;
z_3 = size of the full-time economics faculty, a count;
z_4 = percentage of the economics faculty that are women, proportion (0 to 1);
z_5 = religious affiliation of the college, 0 = no, 1 = yes;
z_6 = percentage of the college faculty that are women, proportion (0 to 1);
$z_7 - z_{10}$ = regional dummy variables, South, Midwest, Northeast, West.

The regressor vectors are

$\mathbf{x}_G = z_1, z_2, z_3, z_4, z_5$ (gender economics course equation),
$\mathbf{x}_W = z_2, z_5, z_6, z_7 - z_{10}$ (women's studies program equation).

TB 17.17

Maximum likelihood estimates of the parameters of Burnett's model were computed by Greene (1998) using her sample of 132 liberal arts colleges; 31 of the schools offer gender economics, 58 have women's studies, and 29 have both. (See Appendix Table F17.1.) The estimated parameters are given in Table 17.17. Both bivariate probit and the single-equation estimates are given. The estimate of ρ is only 0.1359, with a standard error of 1.2359. The Wald statistic for the test of the hypothesis that ρ equals zero is $(0.1359/1.2539)^2 = 0.011753$. For a single restriction, the critical value from the chi-squared table is 3.84, so the hypothesis cannot be rejected. The likelihood ratio statistic for the same hypothesis is $2[-85.6317 - (-85.6458)] = 0.0282$, which leads to the same conclusion. The Lagrange multiplier statistic is 0.003807, which is consistent. This result might seem counterintuitive, given the setting. Surely "gender economics" and "women's studies" are highly correlated, but this finding does not contradict that proposition. The correlation coefficient measures the correlation between the disturbances in the equations, the omitted factors. That is, ρ measures (roughly) the correlation between the outcomes after the influence of the included factors is accounted for. Thus, the value 0.1359 measures the effect after the influence of women's studies is already accounted for. As discussed in the next paragraph, the proposition turns out to be right. The single most important determinant (at least within this model) of whether a gender economics course will be offered is indeed whether the college offers a women's studies program.

The marginal effects in this model are fairly involved, and as before, we can consider several different types. Consider, for example, z_2, academic reputation. There is a direct effect produced by its presence in the gender economics course equation. But there is also an indirect effect. Academic reputation enters the women's studies equation and, therefore, influences the probability that W equals one. Because W appears in the gender economics course equation, this effect is transmitted back to y. The total effect of academic reputation and, likewise, religious affiliation is the sum of these two parts. Consider first the gender economics variable, y. The conditional mean is

$$E[G \mid \mathbf{x}_G, \mathbf{x}_W] = \text{Prob}[W=1]\,E[G \mid W=1, \mathbf{x}_G, \mathbf{x}_W] + \text{Prob}[W=0]\,E[G \mid W=0, \mathbf{x}_G, \mathbf{x}_W]$$

$$= \Phi_2(\mathbf{x}_G'\boldsymbol{\beta}_G + \gamma, \mathbf{x}_W'\boldsymbol{\beta}_W, \rho) + \Phi_2(\mathbf{x}_G'\boldsymbol{\beta}_G, -\mathbf{x}_W'\boldsymbol{\beta}_W, -\rho).$$

Derivatives can be computed using our earlier results. We are also interested in the effect of religious affiliation. Because this variable is binary, simply differentiating the conditional mean function may not produce an accurate result. Instead, we would compute the conditional mean function with this variable set to one and then zero, and take the difference. Finally, what is the effect of the presence of a women's studies program on the probability that the college will offer a gender economics course? To compute this effect, we would compute

$$\text{Prob}[G=1 \mid W=1, \mathbf{x}_G, \mathbf{x}_W] - \text{Prob}[G=1 \mid W=0, \mathbf{x}_G, \mathbf{x}_W].$$

In all cases, standard errors for the estimated marginal effects can be computed using the delta method or the method of Krinsky and Robb.

173
17.18

Table 17.18 presents the estimates of the marginal effects and some descriptive statistics for the data. The calculations were simplified slightly by using the restricted model with $\rho = 0$. Computations of the marginal effects still require the preceding decomposition, but they are simplified by the result that if ρ equals zero, then the bivariate probabilities factor into the products of the marginals. Numerically, the strongest effect appears to be exerted by the representation of women on the faculty; its coefficient of +0.4491 is by far the largest. This variable, however, cannot change by a full unit because it is a proportion. An increase of 1 percent in the presence of women on the faculty raises the probability by only +0.004, which is comparable in scale to the effect of academic reputation. The effect of women on the faculty is likewise fairly small, only 0.0013 per 1 percent change. As might have been expected, the single most important influence is the presence of a women's studies program, which increases the likelihood of a gender economics course by a full 0.1863. Of course, the raw data would have anticipated this result; of the 31 schools that offer a gender economics course, 29 also have a women's studies program and only two do not. Note finally that the effect of religious affiliation (whatever it is) is mostly direct.

17-90

17.17

TABLE 23.13 Estimates of a Recursive Simultaneous Bivariate Probit Model (estimated standard errors in parentheses)

	Single Equation		*Bivariate Probit*	
Variable	*Coefficient*	*Standard Error*	*Coefficient*	*Standard Error*
Gender Economics Equation				
Constant	−1.4176	(0.8768)	−1.1911	(2.2155)
AcRep	−0.01143	(0.003610)	−0.01233	(0.007937)
WomStud	1.1095	(0.4699)	0.8835	(2.2603)
EconFac	0.06730	(0.05687)	0.06769	(0.06952)
PctWecon	2.5391	(0.8997)	2.5636	(1.0144)
Relig	−0.3482	(0.4212)	−0.3741	(0.5264)
Women's Studies Equation				
AcRep	−0.01957	(0.004117)	−0.01939	(0.005704)
PctWfac	1.9429	(0.9001)	1.8914	(0.8714)
Relig	−0.4494	(0.3072)	−0.4584	(0.3403)
South	1.3597	(0.5948)	1.3471	(0.6897)
West	2.3386	(0.6449)	2.3376	(0.8611)
North	1.8867	(0.5927)	1.9009	(0.8495)
Midwest	1.8248	(0.6595)	1.8070	(0.8952)
ρ	0.0000	(0.0000)	0.1359	(1.2539)
ln L	−85.6458		−85.6317	

17.18

TABLE 23.14 Marginal Effects in Gender Economics Model

	Direct	*Indirect*	*Total*	*(Std. Error)*	*(Type of Variable, Mean)*
Gender Economics Equation					
AcRep	−0.002022	−0.001453	−0.003476	(0.001126)	(Continuous, 119.242)
PctWecon	+0.4491		+0.4491	(0.1568)	(Continuous, 0.24787)
EconFac	+0.01190		+0.1190	(0.01292)	(Continuous, 6.74242)
Relig	−0.06327	−0.02306	−0.08632	(0.08220)	(Binary, 0.57576)
WomStud	+0.1863		+0.1863	(0.0868)	(Endogenous, 0.43939)
PctWfac		+0.14434	+0.14434	(0.09051)	(Continuous, 0.35772)
Women's Studies Equation					
AcRep	−0.00780		−0.00780	(0.001654)	(Continuous, 119.242)
PctWfac	+0.77489		+0.77489	(0.3591)	(Continuous, 0.35772)
Relig	−0.17777		−0.17777	(0.11946)	(Binary, 0.57576)

17-91

17.5.6 Endogenous Sampling in a Binary Choice Model

We have encountered several instances of nonrandom sampling in the binary choice setting. In Section 17.3.6, we examined an application in credit scoring in which the balance in the sample of responses of the outcome variable, $C = 1$ for acceptance of an application and $C = 0$ for rejection, is different from the known proportions in the population. The sample was specifically skewed in favor of observations with $C = 1$ to enrich the data set. A second type of nonrandom sampling arose in the analysis of nonresponse/attrition in the GSOEP in Example 17.17. The data suggest that the observed sample is not random with respect to individuals' presence in the sample at different waves of the panel. The first of these represents selection specifically on an observable outcome – the observed dependent variable. We constructed a model for the second of these that relied on an assumption of selection on a set of certain observables – the variables that entered the probability weights. We will now examine a third form of nonrandom sample selection, based crucially on the unobservables in the two equations of a bivariate probit model.

We return to the banking application of Example 17.9. In that application, we examined a binary choice model,

$$\begin{aligned}\text{Prob}(\textit{Cardholder} = 1) &= \text{Prob}(C = 1 \mid \mathbf{x})\\ &= \Phi(\beta_1 + \beta_2\,\textit{Age} + \beta_3\,\textit{Income} + \beta_4\,\textit{OwnRent}\\ &\qquad + \beta_5\,\textit{Months at CurrentAddress}\\ &\qquad + \beta_6\,\textit{SelfEmployed}\\ &\qquad + \beta_7\,\textit{Number of Major Derogatory Reports}\\ &\qquad + \beta_8\,\textit{Number of Minor Derogatory Reports}).\end{aligned}$$

From the point of view of the lender, cardholder status is not the interesting outcome in the credit history, default is. The more interesting equation describes Prob($\textit{Default} = 1 \mid \mathbf{z}, C = 1$). The natural approach, then, would be to construct a binary choice model for the interesting default variable using the historical data for a sample of cardholders. The problem with the approach is that the sample is not randomly drawn – applicants are screened with an eye specifically toward whether or not they seem likely to default. In this application, and in general, there are three economic agents, the credit scorer (e.g., Fair Isaacs), the lender, and the borrower. Each of them has latent characteristics in the equations that determine their behavior. It is these latent characteristics that drive, in part, the application/scoring process and, ultimately, the consumer behavior.

A model that can accommodate these features is (17-$\overset{50}{51}$),

$$\begin{aligned}&S^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad S = 1 \text{ if } S^* > 0, \text{ 0 otherwise},\\ &y^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, \text{ 0 otherwise},\\ &\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix} \Bigg| \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right],\\ &(y, \mathbf{x}_2) \text{ observed only when } S = 1,\end{aligned}$$

which contains an observation rule, $S = 1$, and a behavioral outcome, $y = 0$ or 1. The endogeneity of the sampling rule implies that

$$\text{Prob}(y = 1 \mid S = 1, \mathbf{x}_2) \neq \Phi(\mathbf{x}_2'\boldsymbol{\beta}).$$

17-92

From properties of the bivariate normal distribution, the appropriate probability is

$$\text{Prob}(y = 1 \mid S = 1, \mathbf{x}_1, \mathbf{x}_2) = \Phi\left[\frac{\mathbf{x}_2'\boldsymbol{\beta}_2 + \rho\mathbf{x}_1'\boldsymbol{\beta}_1}{\sqrt{1-\rho^2}}\right].$$

If ρ is not zero, then in using the simple univariate probit model, we are omitting from our model any variables that are in $\mathbf{x}_1$ but not in $\mathbf{x}_2$, and in any case, the estimator is inconsistent by a factor minus $(1 - \rho^2)^{-1/2}$. To underscore the source of the bias, if ρ equals zero, the conditional probability returns to the model that would be estimated with the selected sample. Thus, the bias arises because of the correlation of (i.e., the selection on) the unobservables, ε_1 and ε_2. This model was employed by Wynand and van Praag (1981) in the first application of Heckman's (1979) sample selection model in a nonlinear setting, to insurance purchases, by Boyes, Hoffman and Lowe (1989) in a study of bank lending, by Greene (1992) to the credit card application begun in Example 17.9 and continued in Example 17.22, and hundreds of applications since. [Some discussion appears in Maddala (1983) as well.]

Given that the forms of the probabilities are known, the appropriate log likelihood function for estimation of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and ρ is easily obtained. The log likelihood must be constructed for the joint or the marginal probabilities, not the conditional ones. For the "selected observations," that is, $(y=0,S=1)$ or $(y=1,S=1)$, the relevant probability is simply

minus

$$\text{Prob}(y=0 \text{ or } 1|S=1) \times \text{Prob}(S = 1) = \Phi_2[(2y-1)\mathbf{x}_2'\boldsymbol{\beta}_2, \mathbf{x}_1'\boldsymbol{\beta}_1, (2y-1)\rho]$$

For the observations with $S = 0$, the probability that enters the likelihood function is simply minus $\text{Prob}(S=0|\mathbf{x}_1) = \Phi(-\mathbf{x}_1'\boldsymbol{\beta}_1)$. Estimation is then based on a simpler form of the bivariate probit log likelihood that we examined in Section 17.5.1. Partial effects and post estimation analysis would follow the analysis for the bivariate probit model. The desired partial effects would differ by the application, whether one desires the partial effects from the conditional, joint, or marginal probability would vary. The necessary results are in Section 17.5.3.

Example 17.22 Cardholder Status and Default Behavior

In Example 17.9, we estimated a logit model for cardholder status,

$$\begin{aligned}\text{Prob(Cardholder} = 1) = \text{Prob}(C = 1 \mid \mathbf{x})\\ = \Phi(\beta_1 + \beta_2\,\text{Age} + \beta_3\,\text{Income} + \beta_4\,\text{OwnRent}\\ + \beta_5\,\text{CurrentAddress} + \beta_6\,\text{SelfEmployed}\\ + \beta_7\,\textit{Major Derogatory Reports}\\ + \beta_8\,\textit{Minor Derogatory Reports})\end{aligned}$$

using a sample of 13,444 applications for a credit card. The complication in that example was that the sample was choice based. In the data set, 78.1% of the applicants are cardholders. In the population, at that time, the true proportion was roughly 23.2%, so the sample is substantially choice based on this variable. The sample was deliberately skewed in favor of cardholders for purposes of the original study [Greene (1992)]. The weights to be applied for the WESML estimator are 0.232/0.781 = 0.297 for the observations with $C = 1$ and 0.768/0.219 = 3.507 for observations with $C = 0$. Of the 13,444 applicants in the sample, 10,499 were accepted (given the credit cards). The "default rate" in the sample is 996/10,499 or 9.48%. This is slightly less than the population rate at the time, 10.3%. For purposes of a less complicated numerical example, we will ignore the choice based sampling nature of the data set for the present. An orthodox treatment of both the selection issue and the choice based sampling treatment is left for the exercises [and pursued in Greene (1992).]

percent

percent

17-93

We have formulated the cardholder equation so that it probably resembles the policy of credit scorers, both then and now. A major derogatory report results when a credit account that is being monitored by the credit reporting agency is more than 60 days late in payment. A minor derogatory report is generated when an account is 30 days delinquent. Derogatory reports are a major contributor to credit decisions. Contemporary credit processors such as Fair Isaacs place extremely heavy weight on the "credit score," a single variable that summarizes the credit history and credit-carrying capacity of an individual. We did not have access to credit scores at the time of this study. The selection equation was ~~is~~ given above. The default equation is a behavioral model. There is no obvious standard for this part of the model. We have used three variables, *Dependents*, the number of dependents in the household, *Income*, and *Exp_Income*, which equals the ratio of the average credit card expenditure in the 12 months after the credit card was issued to average monthly income. Default status is measured for the first 12 months after the credit card was issued.

earlier

TB 17.19

Estimation results are presented in Table 17.19. These are broadly consistent with the earlier results – the model with no correlation from Example 17.9 are repeated in Table 17.19. There are two tests we can employ for endogeneity of the selection. The estimate of ρ is 0.41947 with a standard error of 0.11762. The t ratio for the test that ρ equals zero is 3.57, by which we can reject the hypothesis. Alternatively, the likelihood ratio statistic based on the values in Table 17.19 is 2(8670.78831 – 8660.90650) = 19.76362. This is larger than the critical value of 3.84, so the hypothesis of zero correlation is rejected. The results are as might be expected, with one counterintuitive result, that a larger credit burden, expenditure to income ratio, appears to be associated with lower default probabilities, though not significantly so.

Table 17.19 Estimated Joint Cardholder and Default Probability Models

	Endogenous Sample Model		Uncorrelated Equations	
Variable/Equation	**Estimate**	**Standard Error**	**Estimate**	**Standard Error**
		Cardholder Equation		
Constant	0.30516	0.04781 (6.38)	0.31783	0.04790 (6.63)
Age	0.00226	0.00145 (1.56)	0.00184	0.00146 (1.26)
Current Address	0.00091	0.00024 (3.80)	0.00095	0.00024 (3.94)
Own Rent	-0.18758	0.03030 (6.19)	0.18233	0.03048 (5.98)
Income	0.02231	0.00093 (23.87)	0.02237	0.00093 (23.95)
Self Employed	-0.43015	0.05357 (-8.03)	-0.43625	0.05413 (-8.06)
Major Derogatory	-0.69598	0.01871 (-37.20)	-0.69912	0.01839 (-38.01)
Minor Derogatory	-0.04717	0.01825 (-2.58)	-0.04126	0.01829 (-2.26)
		Default Equation		
Constant	-0.96043	0.04728 (-20.32)	-0.81528	0.04104 (-19.86)
Dependents	0.04995	0.01415 (3.53)	0.04993	0.01442 (3.46)
Income	-0.01642	0.00122 (-13.41)	-0.01837	0.00119 (-15.41)
Expend/Income	-0.16918	0.14474 (-1.17)	-0.14172	0.14913 (-0.95)
Correlation	0.41947	0.11762 (3.57)	0.000	0.00000 (0)
log Likelihood	-8660.90650		-8670.78831	

17-94

TABLE 23.15 Binary Choice Fit Measures

Measure	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
LRI	0.573	0.535	0.495	0.407	0.279	0.206	0.000
R^2_{BL}	0.844	0.844	0.823	0.797	0.754	0.718	0.641
λ	0.565	0.560	0.526	0.444	0.319	0.216	0.000
R^2_{EF}	0.561	0.558	0.530	0.475	0.343	0.216	0.000
R^2_{VZ}	0.708	0.707	0.672	0.589	0.447	0.352	0.000
R^2_{MZ}	0.687	0.679	0.628	0.567	0.545	0.329	0.000
Predictions	$\begin{bmatrix} 92 & 9 \\ 5 & 26 \end{bmatrix}$	$\begin{bmatrix} 93 & 8 \\ 5 & 26 \end{bmatrix}$	$\begin{bmatrix} 92 & 9 \\ 8 & 23 \end{bmatrix}$	$\begin{bmatrix} 94 & 7 \\ 8 & 23 \end{bmatrix}$	$\begin{bmatrix} 98 & 3 \\ 16 & 15 \end{bmatrix}$	$\begin{bmatrix} 101 & 0 \\ 31 & 0 \end{bmatrix}$	$\begin{bmatrix} 101 & 0 \\ 31 & 0 \end{bmatrix}$

Before closing this application, we can use this opportunity to examine the fit measures listed in Section 23.4.5. We computed the various fit measures using seven different specifications of the gender economics equation:

1. Single-equation probit estimates, $z_1, z_2, z_3, z_4, z_5, y_2$
2. Bivariate probit model estimates, $z_1, z_2, z_3, z_4, z_5, y_2$
3. Single-equation probit estimates, z_1, z_2, z_3, z_4, z_5
4. Single-equation probit estimates, $z_1, \quad z_3, \quad z_5, y_2$
5. Single-equation probit estimates, $z_1, \quad z_3, \quad z_5$
6. Single-equation probit estimates, $z_1, \qquad z_5$
7. Single-equation probit estimates z_1 (constant only).

The specifications are in descending "quality" because we removed the most statistically significant variables from the model at each step. The values are listed in Table 23.15. The matrix below each column is the table of "hits" and "misses" of the prediction rule $\hat{y} = 1$ if $\hat{P} > 0.5$, 0 otherwise. [Note that by construction, model (7) must predict all ones or all zeros.] The column is the actual count and the row is the prediction. Thus, for model (1), 92 of 101 zeros were predicted correctly, whereas 5 of 31 ones were predicted incorrectly. As one would hope, the fit measures decline as the more significant variables are removed from the model. The Ben-Akiva measure has an obvious flaw in that with only a constant term, the model still obtains a "fit" of 0.641. From the prediction matrices, it is clear that the explanatory power of the model, such as it is, comes from its ability to predict the ones correctly. The poorer the model, the greater the number of correct predictions of $y = 0$. But as this number rises, the number of incorrect predictions rises and the number of correct predictions of $y = 1$ declines. All the fit measures appear to react to this feature to some degree. The Efron and Cramer measures, which are nearly identical, and McFadden's LRI appear to be most sensitive to this, with the remaining two only slightly less consistent.

7
17.5.~~6~~ ~~23.9~~ A MULTIVARIATE PROBIT MODEL

48

In principle, a multivariate probit model would simply extend (17-49 ~~23-44~~) to more than two outcome variables just by adding equations. The resulting equation system, again

17-95

analogous to the seemingly unrelated regressions model, would be

$$\begin{aligned}
&y^*_m = \mathbf{x}'_m\beta_m + \varepsilon_m,\ y_m = 1 \text{ if } y^*_m > 0, 0 \text{ otherwise}, m = 1, \dots, M,\\
&E[\varepsilon_m \mid \mathbf{x}_1, \dots, \mathbf{x}_M] = 0,\\
&\text{Var}[\varepsilon_m \mid \mathbf{x}_1, \dots, \mathbf{x}_M] = 1,\\
&\text{Cov}[\varepsilon_j, \varepsilon_m \mid \mathbf{x}_1, \dots, \mathbf{x}_M] = \rho_{jm},\\
&(\varepsilon_1, \dots, \varepsilon_M) \sim \text{N}_M[\mathbf{0}, \mathbf{R}].
\end{aligned}$$

The joint probabilities of the observed events, $[y_{i1}, y_{i2} \dots, y_{iM} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iM}], i = 1, \dots, n$ that from the basis for the log-likelihood function are the M-variate normal probabilities,

form

$$L_i = \Phi_M(q_{i1}\mathbf{x}'_{i1}\beta_1, \dots, q_{iM}\mathbf{x}'_{iM}\beta_M, \mathbf{R}^*),$$

where

$$\begin{aligned}
q_{im} &= 2y_{im} - 1,\\
\mathbf{R}^*_{jm} &= q_{ij}q_{im}\rho_{jm}.
\end{aligned}$$

The practical obstacle to this extension is the evaluation of the M-variate normal integrals and their derivatives. Some progress has been made on using quadrature for trivariate integration (see Section ~~6.9.6.b~~ 14.9.6.c), but existing results are not sufficient to allow accurate and efficient evaluation for more than two variables in a sample of even moderate size. However, given the speed of modern computers, simulation-based integration using the GHK simulator or simulated likelihood methods (see ~~Chapter 17~~) do allow for estimation of relatively large models. We consider an application in Example ~~23.16~~ 17.23.~~[44]~~ 42

Au: Provide x-ref chap. number ✓

FN 42

The multivariate probit model in another form presents a useful extension of the random effects probit model for panel data (Section ~~23.5.1~~ 17.4.2). If the parameter vectors in all equations are constrained to be equal, we obtain what Bertschek and Lechner (1998) call the "panel probit model,"

$$\begin{aligned}
&y^*_{it} = \mathbf{x}'_{it}\beta + \varepsilon_{it},\ y_{it} = 1 \text{ if } y^*_{it} > 0, 0 \text{ otherwise}, i = 1, \dots, n, t = 1, \dots, T,\\
&(\varepsilon_{i1}, \dots, \varepsilon_{iT}) \sim \text{N}[\mathbf{0}, \mathbf{R}].
\end{aligned}$$

The Butler and Moffitt (1982) approach for this model (see Section ~~23.5.1~~ 17.4.2) has proved useful in many applications. But, their underlying assumption that $\text{Cov}[\varepsilon_{it}, \varepsilon_{is}] = \rho$ is a substantive restriction. By treating this structure as a multivariate probit model with the restriction that the coefficient vector be the same in every period, one can obtain a model with free correlations across periods.~~[45]~~ 43 Hyslop (1999), Bertschek and Lechner (1998), Greene (2004 and Example 23.16), and Cappellari and Jenkins (2006) are applications.

FN 43

17C

1993

[~~44~~] 42 Studies that propose improved methods of simulating probabilities include Pakes and Pollard (1989) and especially Börsch-Supan and Hajivassiliou (1990), Geweke (1989), and Keane (1994). A symposium in the November 1994 issue of *Review of Economics and Statistics* presents discussion of numerous issues in specification and estimation of models based on simulation of probabilities. Applications that employ simulation techniques for evaluation of multivariate normal integrals are now fairly numerous. See, for example, Hyslop (1999) (Example ~~23.7~~ 17.15) which applies the technique to a panel data application with $T = 7$. Example 23.16 17.23 develops a five-variate application.

umlaut

[~~45~~] 43 By assuming the coefficient vectors are the same in all periods, we actually obviate the normalization that the diagonal elements of $\mathbf{R}$ are all equal to one as well. The restriction identifies $T - 1$ relative variances $\rho_{tt} = \sigma^2_T/\sigma^2_T$. This aspect is examined in Greene (2004).

Example ~~23.16~~ 17.23 A Multivariate Probit Model for Product Innovations

Bertschek and Lechner applied the panel probit model to an analysis of the product innovation activity of 1,270 German firms observed in five years, 1984–1988, in response to imports and foreign direct investment. [See Bertschek (1995).] The probit model to be estimated is based on the latent regression

$$y^*_{it} = \beta_1 + \sum_{k=2}^{8} x_{k,it}\beta_k + \varepsilon_{it},\ y_{it} = 1(y^*_{it} > 0),\ i = 1, \dots, 1,270,\ t = 1984, \dots, 1988,$$

where

y_{it} = 1 if a product innovation was realized by firm i in year t, 0 otherwise,
$x_{2,it}$ = Log of industry sales in DM,
$x_{3,it}$ = Import share = ratio of industry imports to (industry sales plus imports),
$x_{4,it}$ = Relative firm size = ratio of employment in business unit to employment in the industry (times 30),
$x_{5,it}$ = FDI share = Ratio of industry foreign direct investment to (industry sales plus imports),
$x_{6,it}$ = Productivity = Ratio of industry value added to industry employment,
$x_{7,it}$ = Raw materials sector = 1 if the firm is in this sector,
$x_{8,it}$ = Investment goods sector = 1 if the firm is in this sector.

The coefficients on import share (β_3) and FDI share (β_5) were of particular interest. The objectives of the study were the empirical investigation of innovation and the methodological development of an estimator that could obviate computing the five-variate normal probabilities necessary for a full maximum likelihood estimation of the model.

TB 17.20 17.20 FN 44

Table ~~23.16~~ 17.20 presents the single-equation, pooled probit model estimates.[~~46~~ 44] Given the structure of the model, the parameter vector could be estimated consistently with any single

TABLE ~~23.16~~ 17.20 Estimated Pooled Probit Model

		Estimated Standard Errors				*Marginal Effects*		
Variable	*Estimate*[a]	*SE(1)*[b]	*SE(2)*[c]	*SE(3)*[d]	*SE(4)*[e]	*Partial*	*Std. Err.*	*t ratio*
Constant	−1.960	0.239	0.377	0.230	0.373	—	—	—
log Sales	0.177	0.0250	0.0375	0.0222	0.0358	0.0683[f]	0.0138	4.96
Rel Size	1.072	0.206	0.306	0.142	0.269	0.413[f]	0.103	4.01
Imports	1.134	0.153	0.246	0.151	0.243	0.437[f]	0.0938	4.66
FDI	2.853	0.467	0.679	0.402	0.642	1.099[f]	0.247	4.44
Prod.	−2.341	1.114	1.300	0.715	1.115	−0.902[f]	0.429	−2.10
Raw Mtl	−0.279	0.0966	0.133	0.0807	0.126	−0.110[g]	0.0503	−2.18
Inv Good	0.188	0.0404	0.0630	0.0392	0.0628	0.0723[g]	0.0241	3.00

[a] Recomputed. Only two digits were reported in the earlier paper.
[b] Obtained from results in Bertschek and Lechner, Table 9.
[c] Based on the Avery et al. (1983) GMM estimator.
[d] Square roots of the diagonals of the negative inverse of the Hessian
[e] Based on the cluster estimator.
[f] Coefficient scaled by the density evaluated at the sample means
[g] Computed as the difference in the fitted probability with the dummy variable equal to one, then zero.

[44] [~~46~~] We are grateful to the authors of this study who have generously loaned us their data for our continued analysis. The data are proprietary and cannot be made publicly available, unlike the other data sets used in our examples.

17-97

period's data, Hence, pooling the observations, which produces a mixture of the estimators, will also be consistent. Given the panel data nature of the data set, however, the conventional standard errors from the pooled estimator are dubious. Because the marginal distribution will produce a consistent estimator of the parameter vector, this is a case in which the cluster estimator (see Section 16.8.4) provides an appropriate asymptotic covariance matrix. Note that the standard errors in column SE(4) of the table are considerably higher than the uncorrected ones in columns 1–3. 14.8.4

The pooled estimator is consistent, so the further development of the estimator is a matter of (1) obtaining a more efficient estimator of β and (2) computing estimates of the cross-period correlation coefficients. ~~The authors proposed a set of GMM estimators based on the orthogonality conditions implied by the single-equation conditional mean functions:~~

$$E\{[y_{it} - \Phi(\mathbf{x}_{it}'\beta)] \mid \mathbf{X}_i\} = 0.$$

~~The orthogonality conditions are~~

$$E\left[\mathbf{A}(\mathbf{X}_i)\begin{pmatrix}[y_{i1} - \Phi(\mathbf{x}_{i1}'\beta)]\\ [y_{i2} - \Phi(\mathbf{x}_{i2}'\beta)]\\ \vdots \\ [y_{iT} - \Phi(\mathbf{x}_{iT}'\beta)]\end{pmatrix}\right] = 0,$$

~~where $\mathbf{A}(\mathbf{X}_i)$ is a $P \times T$ matrix of instrumental variables constructed from the exogenous data for individual i.~~

~~Using only the raw data as $\mathbf{A}(\mathbf{X}_i)$, strong exogeneity of the regressors in every period would provide TK moment equations of the form $E[\mathbf{x}_{it}(y_{is} - \Phi(\mathbf{x}_{is}'\beta))] = 0$ for each pair of periods, or a total of T^2K moment equations altogether for estimation of K parameters in β. [See Wooldridge (1995).] The full set of such orthogonality conditions would be $E[(\mathbf{I}_T \otimes \mathbf{x}_i)\mathbf{u}_i] = \mathbf{0}$, where $\mathbf{x}_i = [\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}']'$, $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})'$ and $u_{it} = y_{it} - \Phi(\mathbf{x}_{it}'\beta)$. This produces 200 orthogonality conditions for the estimation of the 8 parameters. The empirical counterpart to the left-hand side of (6) is~~

$$\mathbf{g}_N(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{A}(\mathbf{X}_i)\begin{pmatrix}[y_{i1} - \Phi(\mathbf{x}_{i1}'\beta)]\\ [y_{i2} - \Phi(\mathbf{x}_{i2}'\beta)]\\ \vdots \\ [y_{iT} - \Phi(\mathbf{x}_{iT}'\beta)]\end{pmatrix}\right].$$

~~The various GMM estimators are the solutions to~~

$$\hat{\beta}_{GMM,A,W} = \arg\min_{\beta}[\mathbf{g}_N(\beta)]'\mathbf{W}[\mathbf{g}_N(\beta)].$$

~~The specific estimator is defined by the choice of instrument matrix $\mathbf{A}(\mathbf{X}_i)$ and weighting matrix, $\mathbf{W}$; the authors suggest several. In their application (see p. 337), only data from period t are used in the tth moment condition. This reduces the number of moment conditions from 200 to 40.~~

AU: Run-in or new ¶?

FN 45

~~As noted~~, The FIML estimates of the model can be computed using the GHK simulator.~~[47]~~ 45 The FIML estimates, ~~Bertschek and Lechner's GMM estimates,~~ and the random effects model using the Butler and Moffit (1982) quadrature method are reported in Table ~~23.17~~ 17.22 21. The ~~FIML and GMM estimates are strikingly similar. This would be expected because both are consistent estimators and the sample is fairly large~~. The correlations reported are based on the FIML estimates. ~~They are not estimated with the GMM estimator. As the authors note, an inefficient estimator of the correlation matrix is available by fitting pairs of equations (years).~~

TB 17.21

AU: Confirm Table 17.21 is correct number

45 ~~[47]~~The full computation required about one hour of computing time. Computation of the single-equation (pooled) estimators required only about 1/100 of the time reported by the authors for the same models, which suggests that the evolution of computing technology may play a significant role in advancing the FIML estimators.

17-98

17.21

TABLE ~~23.17~~ Estimated Constrained Multivariate Probit Model (estimated standard errors in parentheses)

Coefficients	*Full Maximum Likelihood Using GHK Simulator*		*BL GMM*[a]		*Random Effects* $\rho = 0.578$ (0.0189)	
Constant	−1.797**	(0.341)	−1.74**	(0.37)	−2.839	(0.53~~3~~) 4
log Sales	0.154**	(0.0334)	0.15**	(0.034)	0.24~~4~~5	(0.052~~2~~) 3
Relative size	0.953**	(0.160)	0.95**	(0.20)	1.522	(0.25~~1~~) 9
Imports	1.155**	(0.228)	1.14**	(0.24)	1.779	(0.360)
FDI	2.426**	(0.573)	2.59**	(0.59)	3.652	(0.870)
Productivity	−1.578	(1.216)	−1.91*	(0.82)	−2.307	(1.911)
Raw material	−0.292**	(0.130)	−0.28*	(0.12)	−0.477	(0.202)
Investment goods	0.224**	(0.0605)	0.21**	(0.063)	~~0.578~~ 0.331	(0.0~~189~~)
log-likelihood	−3522.85				−3535.55 952	

Estimated Correlations

1984, 1985	0.460**	(0.0301)
1984, 1986	0.599**	(0.0323)
1985, 1986	0.643**	(0.0308)
1984, 1987	0.540**	(0.0308)
1985, 1987	0.546**	(0.0348)
1986, 1987	0.610**	(0.0322)
1984, 1988	0.483**	(0.0364)
1985, 1988	0.446**	(0.0380)
1986, 1988	0.524**	(0.0355)
1987, 1988	0.605**	(0.0325)

Estimated Correlation Matrix

	1984	1985	1986	1987	1988
1984	1.000	(0.658)	(0.599)	(0.540)	(0.483)
1985	0.460	1.000	(0.644)	(0.558)	(0.441)
1986	0.599	0.643	1.000	(0.602)	(0.537)
1987	0.540	0.546	0.610	1.000	(0.621)
1988	0.483	0.446	0.524	0.605	1.000

~~[a]Estimates are BL's WNP joint uniform estimates with $k = 880$. Estimates are from their Table 9, standard errors from their Table 10.~~

*Indicates significant at 95 percent level, ** indicates significant at 99 percent level based on a two-tailed test.

TABLE 23.18 Unrestricted Five-Period Multivariate Probit Model (estimated standard errors in parentheses)

Coefficients	*1984*	*1985*	*1986*	*1987*	*1988*	*Constrained*
Constant	−1.802** (0.532)	−2.080** (0.519)	−2.630** (0.542)	−1.721** (0.534)	−1.729** (0.523)	−1.797** (0.341)
log Sales	0.167** (0.0538)	0.178** (0.0565)	0.274** (0.0584)	0.163** (0.0560)	0.130** (0.0519)	0.154** (0.0334)
Relative size	0.658** (0.323)	1.280** (0.330)	1.739** (0.431)	1.085** (0.351)	0.826** (0.263)	0.953** (0.160)
Imports	1.118** (0.377)	0.923** (0.361)	0.936** (0.370)	1.091** (0.338)	1.301** (0.342)	1.155** (0.228)
FDI	2.070** (0.835)	1.509* (0.769)	3.759** (0.990)	3.718** (1.214)	3.834** (1.106)	2.426** (0.573)
Productivity	−2.615 (4.110)	−0.252 (3.802)	−5.565 (3.537)	−3.905 (3.188)	−0.981 (2.057)	−1.578 (1.216)
Raw material	−0.346 (0.283)	−0.357 (0.247)	−0.260 (0.299)	0.0261 (0.288)	−0.294 (0.218)	−0.292 (0.130)
Investment goods	0.239** (0.0864)	0.177* (0.0875)	0.0467 (0.0891)	0.218* (0.0955)	0.280** (0.0923)	0.224** (0.0605)

17-99

final 21

FINAL ESTIMATES 17.22

~~as bivariate probit models.)~~ Also noteworthy in Table ~~23.17~~ is the divergence of the random effects estimates from the ~~other two sets~~. The log-likelihood function is −3535.55 for the random effects model and −3522.85 for the unrestricted model. The chi-squared statistic for the 9 restrictions of the equicorrelation model is 25.4. The critical value from the chi-squared table for 9 degrees of freedom is 16.9 for 95 percent and 21.7 for 99 percent significance, so the hypothesis of the random effects model would be rejected.

nine//

~~Table 23.16 reports the coefficients of a fully unrestricted model that allows the coefficients~~ to vary across the periods. (The correlations in parentheses in Table 23.17 are computed using this model.) There is a surprising amount of variation in the parameter vector. The log-likelihood for the full unrestricted model is −3494.57. The chi-squared statistic for testing the 32 restrictions of the homogeneity hypothesis is twice the difference, or 56.56. The critical value from the chi-squared table with 32 degrees of freedom is 46.19, so the hypothesis of homogeneity would be rejected statistically. It does seem questionable whether, in theoretical terms, the relationship estimated in this application should be this volatile, ~~however~~.

23.10 ANALYSIS OF ORDERED CHOICES

Some multinomial-choice variables are inherently ordered. Examples that have appeared in the literature include the following:

1. Bond ratings
2. Results of taste tests
3. Opinion surveys
4. The assignment of military personnel to job classifications by skill level
5. Voting outcomes on certain programs
6. The level of insurance coverage taken by a consumer: none, part, or full
7. Employment: unemployed, part time, or full time

In each of these cases, although the outcome is discrete, the multinomial logit or probit model would fail to account for the ordinal nature of the dependent variable.[48] Ordinary regression analysis would err in the opposite direction, however. Take the outcome of an opinion survey. If the responses are coded 0, 1, 2, 3, or 4, then linear regression would treat the difference between a 4 and a 3 the same as that between a 3 and a 2, whereas in fact they are only a ranking.

23.10.1 THE ORDERED PROBIT MODEL

The ordered probit and logit models have come into fairly wide use as a framework for analyzing such responses [Zavoina and McElvey (1975)]. The model is built around a latent regression in the same manner as the binomial probit model. We begin with

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

[48] In two papers, Beggs, Cardell, and Hausman (1981) and Hausman and Ruud (1986), the authors analyze a richer specification of the logit model when respondents provide their rankings of the full set of alternatives in addition to the identity of the most preferred choice. This application falls somewhere between the conditional logit model and the ones we shall discuss here in that, rather than provide a single choice among J either unordered or ordered alternatives, the consumer chooses one of the $J!$ possible orderings of the set of unordered alternatives.

17.6 SUMMARY AND CONCLUSIONS

This chapter has surveyed a large range of techniques for modeling a binary choice variable. The model for choice between two outcomes provides the framework for a large proportion of the analysis of microeconomic data. Thus, we have given a very large amount of space to this model in its own right. In addition, many issues in model specification and estimation that appear in more elaborate settings, such as those we will examine in the next chapter, can be formulated as extensions of the binary choice model of this chapter. Binary choice modeling provides a convenient point to study endogeneity in a nonlinear model, issues of nonresponse in panel data sets, and general problems of estimation and inference with longitudinal data. The binary probit model in particular has provided the laboratory case for theoretical econometricians such as those who have developed methods of bias reduction for the fixed effects estimator in dynamic nonlinear models.

We began the analysis with the fundamental parametric probit and logit models for binary choice. Estimation and inference issues such as the computation of appropriate covariance matrices for estimators and partial effects are considered here. We then examined familiar issues in modeling, including goodness of fit and specification issues such as the distributional assumption, heteroscedasticity and missing variables. As in other modeling settings, endogeneity of some right hand variables presents a substantial complication in the estimation and use of nonlinear models such as the probit model. We examined the problem of endogenous right hand side variables, and in two applications, problems of endogenous sampling. The analysis of binary choice with panel data provides a setting to examine a large range of issues that reappear in other applications. We reconsidered the familiar pooled, fixed and random effects estimator estimators, and found that much of the wisdom obtained in the linear case does not carry over to the nonlinear case. The incidental parameters problem, in particular, motivates a considerable amount of effort to reconstruct the estimators of binary choice models. Finally, we considered some multivariate extensions of the probit model. As before, the models are useful in their own right. Once again, they also provide a convenient setting in which to examine broader issues, such as more detailed models of endogeneity nonrandom sampling, and computation requiring simulation.

Chapter 18 will continue the analysis of discrete choice models with three frameworks; unordered multinomial choice, ordered choice and models for count data. Most of the estimation and specification issues we have examined in this chapter will reappear in these settings.

17-161

AO: Terms with blue checkmarks were not bold KTs in chapter

Key Terms and Concepts

- Attributes
- Attrition bias
- Average partial effect
- Binary choice model
- ✓ Bivariate probit
- ✓ Bootstrapping
- Butler and Moffitt method
- Characteristics
- Choice-based sampling
- Chow test
- Complementary log log model
- Conditional likelihood function
- Control function
- Event count
- Fixed effects model
- Generalized residual
- ✓ Goodness of fit measure
- Gumbel model
- Heterogeneity
- Heteroscedasticity
- Incidental parameters problem
- Index function model
- Initial conditions
- Interaction effect
- Inverse probability weighted {IPW}
- Lagrange multiplier test
- Latent regression
- Likelihood equations
- Likelihood ratio test
- Linear probability model
- Logit
- Marginal effects
- Maximum likelihood
- Maximum simulated likelihood {MSL}
- Method of scoring
- Minimal sufficient statistic
- Multinomial choice
- Multivariate probit model
- Nonresponse bias
- Ordered choice model
- Persistence
- Probit
- Quadrature
- Qualitative response {QR}
- Quasi-~~MLE~~ =maximum likelihood estimator (QMLE)
- Random effects model
- ✓ Random parameters logit model
- Random utility model

→

17-102

- Recursive model
- Robust covariance estimation
- Sample selection bias
- State dependence
- Tetrachoric correlation
- Unbalanced sample

Multinomial choice
Event count
Average partial effect
Interaction effect
Nonresponse bias

Attrition bias
Sample Selection bias
Selection on unobservables
Inverse probability weight
Complementary log log model

17-103

- ~~Nested logit model~~
- ~~Nonnested models~~
- ~~Normit~~
- ✓ Ordered choice ~~model~~
- ✓ Persistence
- ✓ Probit
- ✓ Quadrature
- ~~Qualitative choice~~
- ✓ Qualitative response
- ✓ Quasi-MLE
- ~~Random coefficients~~
- ✓ Random effects model
- ~~Random parameters logit model~~
- ✓ Random utility ~~model~~
- ~~Ranking~~
- ✓ Recursive model
- ~~Revealed preference data~~
- ✓ Robust covariance estimation
- ~~Semiparametric estimation~~
- ✓ State dependence
- ~~Stated choice data~~
- ~~Stated choice experiment~~
- ✓ Tetrachoric correlation
- ✓ Unbalanced sample
- ~~Unordered choice~~

Exercises

1. A binomial probability model is to be based on the following index function model:

$$\begin{aligned} y^* &= \alpha + \beta d + \varepsilon, \\ y &= 1, \quad \text{if } y^* > 0, \\ y &= 0 \quad \text{otherwise.} \end{aligned}$$

The only regressor, d, is a dummy variable. The data consist of 100 observations that have the following:

		y = 0	y = 1
d	0	24	28
	1	32	16

Obtain the maximum likelihood estimators of α and β, and estimate the asymptotic standard errors of your estimates. Test the hypothesis that β equals zero by using a Wald test (asymptotic t test) and a likelihood ratio test. Use the probit model and then repeat, using the logit model. Do your results change? (Hint: Formulate the log-likelihood in terms of α and $\delta = \alpha + \beta$.)

2. Suppose that a linear probability model is to be fit to a set of observations on a dependent variable y that takes values zero and one, and a single regressor x that varies continuously across observations. Obtain the exact expressions for the least squares slope in the regression in terms of the mean(s) and variance of x, and interpret the result.

3. Given the data set

y	1	0	0	1	1	0	0	1	1	1
x	9	2	5	4	6	7	3	5	2	6'

estimate a probit model and test the hypothesis that x is not influential in determining the probability that y equals one.

4. Construct the Lagrange multiplier statistic for testing the hypothesis that all the slopes (but not the constant term) equal zero in the binomial logit model. Prove that the Lagrange multiplier statistic is nR^2 in the regression of $(y_i = p)$ on the x's, where p is the sample proportion of 1's.

17-104

move to 8

~~5. We are interested in the ordered probit model. Our data consist of 250 observations, of which the response are~~

y	0	1	2	3	4
n	50	40	45	80	35

~~Using the preceding data, obtain maximum likelihood estimates of the unknown parameters of the model. (Hint: Consider the probabilities as the unknown parameters.)~~

5 ~~6~~. The following hypothetical data give the participation rates in a particular type of recycling program and the number of trucks purchased for collection by 10 towns in a small mid-Atlantic state:

Town	1	2	3	4	5	6	7	8	9	10
Trucks	160	250	170	365	210	206	203	305	270	340
Participation%	11	74	8	87	62	83	48	84	71	79

The town of Eleven is contemplating initiating a recycling program but wishes to achieve a 95 percent rate of participation. Using a probit model for your analysis,

a. How many trucks would the town expect to have to purchase to achieve its goal? (Hint: You can form the log likelihood by replacing y_i with the participation rate (e.g., 0.11 for observation 1) and $(1 - y_i)$ with 1—the rate in (~~23-33~~). 17-23 22

b. If trucks cost \$20,000 each, then is a goal of 90 percent reachable within a budget of \$6.5 million? (That is, should they *expect* to reach the goal?)

c. According to your model, what is the marginal value of the 301st truck in terms of the increase in the percentage participation?

6 ~~7~~. A data set consists of $n = n_1 + n_2 + n_3$ observations on y and x. For the first n_1 observations, $y = 1$ and $x = 1$. For the next n_2 observations, $y = 0$ and $x = 1$. For the last n_3 observations, $y = 0$ and $x = 0$. Prove that neither (~~23-19~~) nor (~~23-21~~) has a solution.
17-19 18 17-21 20

7 ~~8~~. Prove (~~23-28~~). 17-30

8 ~~9~~. In the panel data models estimated in Section ~~23.5~~ 17.4, neither the logit nor the probit model provides a framework for applying a Hausman test to determine whether fixed or random effects is preferred. Explain. (Hint: Unlike our application in the linear model, the incidental parameters problem persists here.)

Applications

1. Appendix Table F~~23~~17.1 provides Fair's (1978) *Redbook* survey on extramarital affairs. The data are described in Application 1 at the end of Chapter ~~24~~ 18 and in Appendix F. The variables in the data set are as follows:

id = an identification number,
C = constant, value = 1,
yrb = a constructed measure of time spent in extramarital affairs,

17-105

End 17

$v1$ = a rating of the marriage, coded 1 to 4,
$v2$ = age, in years, aggregated,
$v3$ = number of years married,
$v4$ = number of children, top coded at 5,
$v5$ = religiosity, 1 to 4, 1 = not, 4 = very,
$v6$ = education, coded 9, 12, 14, 16, 17, 20,
$v7$ = occupation,
$v8$ = husband's occupation,

and three other variables that are not used. The sample contains a survey of 6,366 married women, conducted by *Redbook* magazine. For this exercise, we will analyze, first, the binary variable

$$A = 1 \text{ if } yrb > 0, 0 \text{ otherwise.}$$

The regressors of interest are $v1$ to $v8$; however, not necessarily all of them belong in your model. Use these data to build a binary choice model for A. Report all computed results for the model. Compute the marginal effects for the variables you choose. Compare the results you obtain for a probit model to those for a logit model. Are there any substantial differences in the results for the two models?

move to 18

~~2. Continuing the analysis of the first application, we now consider the self-reported rating, $v1$. This is a natural candidate for an ordered choice model, because the simple four-item coding is a censored version of what would be a continuous scale on some subjective satisfaction variable. Analyze this variable using an ordered probit model. What variables appear to explain the response to this survey question? (Note, the variable is coded 1, 2, 3, 4. Some programs accept data for ordered choice modeling in this form, e.g., *Stata*, while others require the variable to be coded 0, 1, 2, 3, e.g., *LIMDEP*. Be sure to determine which is appropriate for the program you are using and transform the data if necessary.) Can you obtain the partial effects for your model? Report them as well. What do they suggest about the impact of the different independent variables on the reported ratings?~~