The resulting likelihood function is

$$\mathrm{L}(\boldsymbol{\beta}, \lambda_1, \ldots, \lambda_A) = \prod_{\mathrm{i}=1}^{N} \left[\prod_{s=1}^{a_i - 1} (1 - F\left(\lambda_s + \mathbf{x}_i'(t_{s-1})\boldsymbol{\beta}\right)) \right] \times F\left(\lambda_{a_i} + \mathbf{x}'(t_{a_i - 1})\boldsymbol{\beta}\right).$$

This is similar to (17.42), the log-likelihood for discrete time PH model, aside from the choice of function F. The hazard (17.40) is the extreme value cdf evaluated at $\ln \lambda_{0a} + \mathbf{x}(t_{a-1})'\boldsymbol{\beta}$, so (17.40) yields the complementary log-log model binary choice model (see Table 14.3) rather than the more commonly used logit or probit model.

17.11. Duration Example: Unemployment Duration

The following empirical application uses the data of McCall (1996), generously provided to us by the author Brian McCall. The data set is derived from the January Current Population Survey's Displaced Workers Supplements (DWS) for the years 1986, 1988, 1990, and 1992. We refer to the duration measure (spell) in this example as unemployment duration, though more accurately it represents joblessness duration since DWS does not provide information as to whether a person is looking for job or not.

For this application, information on the part-time or full-time status of the first postdisplacement job is required. To determine whether the first postdisplacement job was part-time or full-time, the following method is adopted. The first postdisplacement job is designated as part-time if a subject was still in that job at the time of the survey and if the subject was working less than 35 hours per week in that job in the previous week.

Table 17.6 defines the key economic covariates used to explain joblessness duration. The number of covariates in the models estimated is quite large, but in the interest of brevity only a subset is listed. McCall (1996) provides a fuller description.

Table 17.6. *Unemployment Duration: Description of Variables*

Variable Name	Variable Label	Mean
spell	periods jobless: two-week interval	6.248
CENSOR1	1 if reemployed at full-time job	0.321
CENSOR2	1 if reemployed at part-time job	0.102
CENSOR3	1 if reemployed but left job: pt–ft status unknown	0.172
CENSOR4	1 if still jobless	0.375
UI	1 if filed UI claim	0.553
RR	eligible replacement rate	0.454
DR	eligible disregard rate	0.109
TENURE	tenure years in lost job	4.114
LOGWAGE	log weekly earnings	5.693

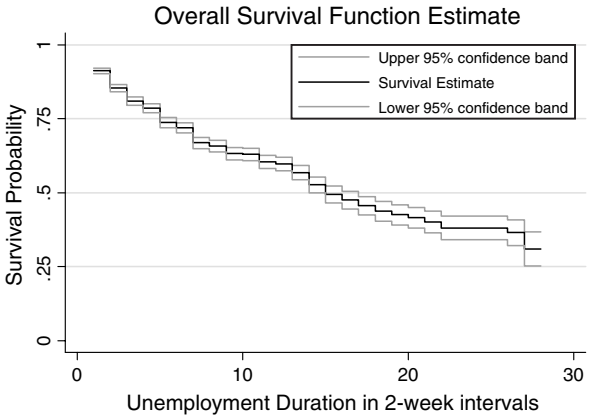

Figure 17.3: Unemployment duration: Kaplan-Meier estimate of survival function. U.S. data from 1986–92 on 3343 spells, some incomplete.

Unemployment durations have been measured in two-week intervals. Four binary variables (CENSOR1, CENSOR2, CENSOR3, and CENSOR4) have been introduced to indicate the status of the first postdisplacement job. For the analysis in this chapter we use CENSOR1. Thus a spell is complete if person is re-employed at a full-time job. Another indicator variable UI is used to denote whether the subject filed an unemployment claim or not. Replacement rate, which is the weekly benefit amount divided by the amount of weekly earnings in the lost job, is represented by the variable RR. "Disregard" is defined to be the threshold amount up to which recipients of unemployment insurance who accept part-time work can earn without any reduction in unemployment benefits. Disregard rate is the disregard divided by weekly earnings in the lost job. It is described by the variable DR in this example. As we can see, all the other variables are self-explanatory.

We begin with a descriptive analysis of the duration data. The simplest first step is to plot the Kaplan–Meier survival curve, which is shown in Figure 17.3 by the dark line. The lighter lines around the estimated Kaplan–Meier survival curve represent 95% confidence intervals developed in Section 17.5.2. As expected, the estimated survival curve declines rapidly at first and then slowly.

As we see from Table 17.7, after the first period the survival probability is 0.91, indicating that roughly 9% of the sampled individuals have terminated their spell within the first two weeks of beginning joblessness spell.

In Figure 17.4, we plot the survival function by UI, that is, by whether the subject claims unemployment insurance or not. Again, as one can expect, it shows that those who claim unemployment insurance are more likely to remain unemployed than those who do not claim unemployment insurance.

The Nelson–Aalen cumulative hazard in Figure 17.5 shows little variation in the hazard rate, which translates into an approximately linear hazard. If the crude hazard rate varies a lot, then the cumulative hazard would appear nonlinear.

Table 17.7. *Unemployment Duration: Kaplan–Meier Survival and Nelsen–Aalen Cumulated Hazard Functions*

Time	Survivor Function	Cumulative Hazard
1	0.9121	0.0879
2	0.8541	0.1514
3	0.8103	0.2027
4	0.7864	0.2322
5	0.7376	0.2943
⋮	⋮	⋮
12	0.5974	0.5005
13	0.5680	0.5496
14	0.5270	0.6219
⋮	⋮	⋮
26	0.3651	0.9809
27	0.3098	1.1325
28	0.3098	1.1325

The cumulated hazard functions by UI recipiency, shown in Figure 17.6, exhibit the expected pattern: The hazard increases at a higher rate for those who do not claim unemployment insurance than it does for those who do.

Next we consider four parametric regression models using the covariates UI, RR, DR, and LOGWAGE and the interaction terms RRUI and DRUI. The four models are exponential, Weibull, Gompertz, and Cox PH. Writing the hazard function as

$$\lambda(t|\mathbf{x}) = \lambda_0(t,\alpha)\phi(\mathbf{x},\beta) = \lambda_0(t,\alpha)\exp(\mathbf{x}'\beta),$$

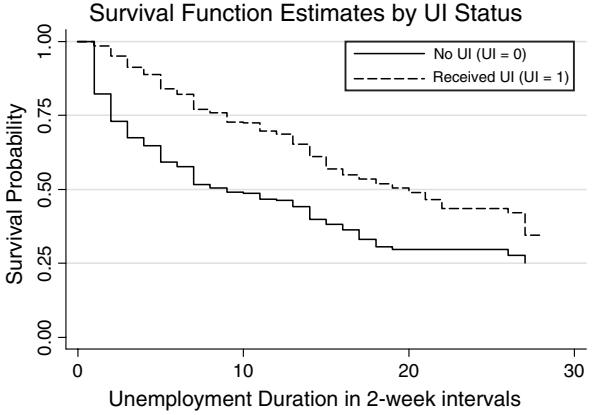

Figure 17.4: Unemployment duration: estimated survival functions by whether or not subjects receive unemployment insurance. Same data as Figure 17.3.

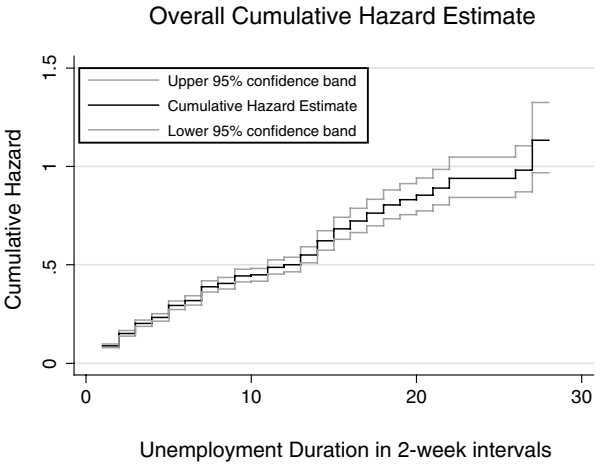

Figure 17.5: Unemployment duration: Nelson-Aalen estimate of cumulative hazard function. Same data as Figure 17.3.

recall that exponential hazard assumes $\lambda_0(t, \alpha) = \text{constant} = \exp(a)$ for some constant a, the Weibull model assumes $\lambda_0(t, \alpha) = \exp(a)\alpha t^{\alpha-1}$ (i.e., monotonic hazards), Gompertz assumes $\lambda_0(t, \alpha) = \exp(a)\exp(\gamma t)$, and the Cox PH model has no intercept and makes no assumption about the shape of the baseline hazard. Recall also that the formulation here is of the proportional hazard type and can also be interpreted either as a parametric regression model or as an AFT model. In this parameterization of the likelihood function, the parameters (α, β) are estimated. These are given in Table 17.8 with the associated t-statistics. We also list the negative of the log-likelihood, but recall that for the Cox PH model it is the partial log-likelihood. Both exponential and Gompertz models fit equally well. The Weibull model provides the

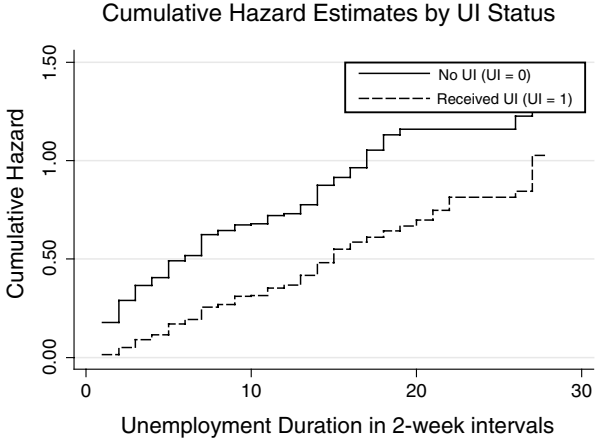

Figure 17.6: Unemployment duration: estimated cumulative hazard functions by whether or not receive unemployment insurance. Same data as Figure 17.3.

Table 17.8. *Unemployment Duration: Estimated Parameters from Four Parametric Models*

	Exponential		Weibull		Gompertz		Cox PH	
Var	**coeff.**	***t***	**coeff.**	***t***	**coeff.**	***t***	**coeff.**	***t***
RR	0.472	0.79	0.448	0.70	0.472	0.78	0.522	0.91
DR	−0.576	−0.75	−0.427	−0.53	−0.563	−0.74	−0.753	−1.04
UI	−1.425	−5.71	−1.496	−5.67	−1.428	−5.69	−1.317	−5.55
RRUI	0.966	0.92	1.105	1.57	0.969	1.58	0.882	1.52
DRUI	−0.199	−0.20	−0.299	−0.28	−0.211	−0.21	−0.095	−0.10
LOGWAGE	0.35	3.03	0.37	2.99	0.35	3.03	0.34	3.03
CONS	−4.079	−4.65	−4.358	−4.74	−4.097	−4.65	–	–
α			1.129					
$-\ln L$	2700.7		2687.6		2700.6		–	

best fit. As we see from Table 17.8, the fit of the Weibull model exhibits positive state dependence ($\alpha = 1.129 > 1$); that is, the probability of the spell terminating increases as the spell lengthens.

For all the models considered, only UI and LOGWAGE are significant whereas other covariates are not. The estimated coefficient of UI is negative for all models, implying that the joblessness spell of those who claim unemployment insurance terminates slower. There is little variation of the estimates of UI across different models: This estimate in Weibull and Gompertz models is approximately 5% and 0.2% higher in absolute value than that in the exponential model, whereas it is 8% lower in the Cox PH model. Similarly, the estimate of the coefficient of LOGWAGE is positive for all the models and exhibits very little variation across models.

Whereas in the econometric literature it is common to report the estimate of (α, β) coefficients of the hazard function in AFT metric, in the biostatistics literature a different parameterization is often used based on the PH metric. Note that the hazard ratio $\lambda(t|\mathbf{x})/\lambda_0(t, \alpha) = \phi(\mathbf{x}, \beta) = \exp(\mathbf{x}'\beta)$. For a categorical 0/1 scalar variable x, the impact of a change from 0 to 1 is given by $\exp(\beta) - 1$, which measures impact relative to the baseline hazard. Numerous packages give the users an option to estimate the model in either or both metrics. The relative merits of the two parameterization are discussed in Cleves, Gould, and Guitirrez (2002).

Consider the exponential specification in Table 17.9 where the coefficients are exponentials of the corresponding ones Table 17.8. Here UI has hazard ratio 0.241. This means that belonging to the category of subjects that claims unemployment insurance decreases the hazard by nearly 76% over the baseline hazard. Similarly, for Weibull, Gompertz, and Cox PH models, the hazard decreases by about 78%, 76%, and 73%, respectively.

For this example, we have taken into account right-censoring and have ignored the role of unobserved heterogeneity. Hence the results obtained from the three models are qualitatively similar. However, the relatively few included variables with significant

Table 17.9. *Unemployment Duration: Estimated Hazard Ratios from Four Parametric Models*

	Exponential		Weibull		Gompertz		Cox PH	
Var	β	t	β	t	β	t	β	t
RR	1.603	0.63	1.565	0.57	1.604	0.62	1.686	0.71
DR	0.562	−1.02	0.653	−0.66	0.570	−0.99	0.471	−1.55
UI	0.241	−12.65	0.224	−13.12	0.240	−12.65	0.268	−11.53
RRUI	2.626	1.01	2.760	0.99	2.635	1.01	2.416	1.01
DRUI	0.819	−0.22	0.742	−0.33	0.810	−0.23	0.909	−0.10
LOGWAGE	1.420	2.56	1.441	0.08	1.42	2.55	1.40	2.57
α			1.129					
$-\ln L$	2700.7		2687.6		2700.6		–	

coefficients probably indicates that large unexplained variation (perhaps caused by unobserved heterogeneity) may be a serious problem. This issue is considered further in the next chapter.

17.12. Practical Considerations

Most computer packages offer a good selection of computer programs for parametric survival analysis. Standard nonparametric Kaplan–Meier survival function estimates, with or without confidence intervals, with both numeric and graphic output are widely available. In some cases survival analysis modules are sufficiently detailed to warrant a special manual. For example, Allison (1995) offers a practical guide to survival analysis in the SAS system; Cleves et al. (2002) provide a tutorial style guide to survival analysis in STATA. Not only do these guides explain the mechanics of implementing particular program commands, but in many cases they provide insightful expositions of the subtleties arising from specific features of data, alternative parameterizations, and interpretation of results. A convenient way to learn about duration data analysis is by using the examples in econometrics or statistical packages such as LIMDEP, STATA, SAS, or S-Plus. The program manuals are themselves excellent sources of information for standard models.

17.13. Bibliographic Notes

17.3–17.7 Kalbfleisch and Prentice (1980, 2002) is the classic statistical reference for survival analysis, with emphasis on the Cox model. Other useful sources include Lawless (1982) and Cox and Oakes (1984) and the considerable number of statistical texts on survival analysis that now exist. For a Bayesian treatment see Ibrahim, Chen, and Sinha (2001). Recent statistical work has increasingly emphasized the counting process approach, detailed in Fleming and Harrington (1991) and Andersen et al. (1993).