Errata for the First Printing of Survival Analysis Using S: Analysis of Time-to-Event Data

Tableman & Kim

February 3, 2005

See also “What’s additional and/or new since the first printing?” on the book’s website.

Key: L2 means two lines down from the top of page or the place we reference. Similarly, L-2 means two lines up from the bottom of page or the place we reference.

p. 2: L2 from **Example 2. CNS lymphoma data:** Health & Science University

p. 5: Expression (1.2) $S(t) = P(T > t) = \dots$

p. 7: h(t) on the seesaw graph should be H(t).

p. 11: L7: T_{15} should be $T_{(15)}$.

p. 21: L2 from the first S command: Insert the following right after “use `sep = " "`.”:

For tabbed data entry, use `sep = "\t"`.

p. 26:

```
aml1 <- aml[aml$group==1,] # Maintained group only

esf.fit <- survfit(Surv(aml1$weeks,status) ~ 1)

 # lab=c(x,y,llen) desired number of tick intervals on the x and y
 # axes and the length of labels on both axes. The default is
 # c(5,5,7).

> mtext("The Empirical Survivor Function of the AML Data",
    side=3,line=-3) # side: integer indicating the side on which
 # text is to be placed. 1 is the bottom, 2 is the left, 3 is the
 # top and 4 is the right. line: distance of text from the plot
 # (measured out from the plot in units of mex sized character
 # heights). If line=0 and outer=F, the text is printed just outside
 # the plot area; positive values put the text farther from the plot,
 # negative values closer. mex: states which font is to be used to
 # measure the margins.
```

p. 32: In the middle:

$$\text{s.e.}\left(\widehat{S}(13)\right) \quad = \quad \sqrt{.0136} = .1166$$

p. 35: L-3 from **Note:**: expected survival time $\widehat{\text{mean}}$ = 52.6 weeks with s.e.($\widehat{\text{mean}}$) = 19.8 weeks.

p. 37: L1 from expression (2.16): … cumulative hazard function (2.7) with s.e. being the

p. 38:

L2: Insert "See Remark 7." at the end.

The last term in expression (2.21) should have a negative sign instead of a positive sign.

Right before item 7: Insert "See Remark 7."

p. 41: L-5 and L-4: $\chi_{(1)}$ should be $\chi^2_{(1)}$.

p. 45: L2 from **Hazard ratio as a measure of effect:** $\tilde{h}_i$ should be $\tilde{h}(t_i)$.

p. 49: L2 from the beginning of the last paragraph: … those who were not traced had medical aid.

p. 53: Within **Tips:**

```
> fit.sex <- survfit(Surv(lzeit,tod)~diab+strata(sex),data=diabetes)
```

p. 60: L3: $-\infty < Z < \infty$ should be $-\infty < z < \infty$.

p. 74:

Table 3.2: [28.82,126.76] should be [28.81,126.76].

In the middle:

$$\log\left(-\log(\widehat{S}(t))\right) \quad \overset{a}{\sim} \quad N\left(\log(-\log(S(t))) = \log(\lambda t), \frac{1}{n_u}\right).$$

p. 76: `> C.I.median` should be `> C.I.median1`

p. 80: Figure 3.9 caption and title: AML data should be AML1 data.

p. 86: Table under second bullet from **Results:** Header for second column should be "max log-likeli" and "log $\big(L(\widehat{\theta},\widehat{\beta}^*)\big)$".

p. 88:

Table: $\widehat{\sigma}$ should be $\widehat{\alpha}$.

L-7: Insert "follow-up" between the words "over" and "time".

L-6: Delete "The graph of $h(t|1)$ and $h(t|0)$ is a set of parallel lines over time." Replace with: That is, the graph of the HR over follow-up time is a horizontal line with height $\exp(\beta)$.

p. 93: Exercise 3.2 (a):

Tips:

```
> diabetes1 <- diabetes[diabets$diab==1,]
> weib.fit <- survReg(Surv(lzeit,tod)~1,dist="weib",data=diabetes1)
> alphahat <- 1/weib.fit$scale
> lzeit.u <- lzeit[tod==1]
> nu <- length(lzeit.u)
> muhat <- weib.fit$coeff
> scalehat <- rep(exp(muhat),nu)
> lzeit.u <- sort(lzeit.u)
> Shat <- 1 - pweibull(lzeit.u,alphahat,scalehat)
```

p. 95: Fourth bullet under **Example 1.** should be: $T =$ survival time.

p. 103: L-6: Then it follows that, if $g(\cdot)$ is an increasing function,

p. 106: The first bullet:

$$S(t|\underline{x}) = S_0^*\big(\exp(-\underline{x}'\underline{\beta}^*)t\big) = \frac{1}{1+\big(\exp(y-\beta_0^*-\underline{x}'\underline{\beta}^*)\big)^{\frac{1}{\sigma}}}\ ,$$

p. 107: L4: log-likelihood

p. 108: L-3 from **The S code for computing the AIC for a number of specified distributions**:

$$\begin{aligned} \text{intercept only: } Y &= \log(T) = \beta_0^* + \sigma Z \\ \text{both: } Y &= \log(T) = \beta_0^* + \beta_1^* x + \sigma Z, \end{aligned}$$

p. 116: Table under second bullet: Fifth entry in Estimate column should be 32626.72

p. 118: Exercise 4.1 (b) **Tips:**

"lzet" in `> fit.b <- survReg(Surv(lzet,tod) ...` should be "lzeit".

p. 119:

L6: Table 3.2 in Chapter 3

Exercise 4.3 (b): Insert "at arbitrary two different ages" at the end.

p. 123: L3 from the output: Replace "Recall that the hazard (risk) function," by "Recall from expression (1.5), page 6,"

p. 124: L-4 from **Example:** "If the scope argument ..." should be "If the direction argument ..."

p. 125: AGE60 = 1 if AGE $\geq$ 60 should be AGE60 = 1 if AGE $\leq$ 60

p. 132:

The model

```
> cox.fit <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE+GROUP+
                                 SEX+strata(AGE60))
```

was used to construct Figure 5.3. This is wrong for our purpose of displaying the interaction effect between SEX and AGE60 on survival after adjusting for the GROUP and KPS.PRE. effects. Hence, delete the material from L8 through L17. That is, delete "`> cox.fit` ... for the three quartiles." Replace with the following:

```
> cox.fit <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP+
                               strata(factor(SEX),factor(AGE60)))
```

which adjusts for the GROUP and KPS.PRE. effects. We then set GROUP = 1 and KPS.PRE. = 80 and obtain the summary of the adjusted quantiles and means in the same four strata (cells) as in Figure 5.2 using `survfit` as follows:

```
> survfit(cox.fit,data.frame(GROUP=1,KPS.PRE.=80))
> summary(survfit(cox.fit,data.frame(GROUP=1,KPS.PRE.=80)))
```

Figure 5.3 displays both ordinal and disordinal interactions. The survival curve for the females who are younger than 60 years never steps below 0.50 (see `> summary` above). In order to produce the median plot, we set the median survival time since 1st BBBD for this stratum at 1.375 years, which is the .632-quantile.

p. 133: Figure 5.3 is wrong. Replace it with Figure 5.3 printed in this errata sheet.

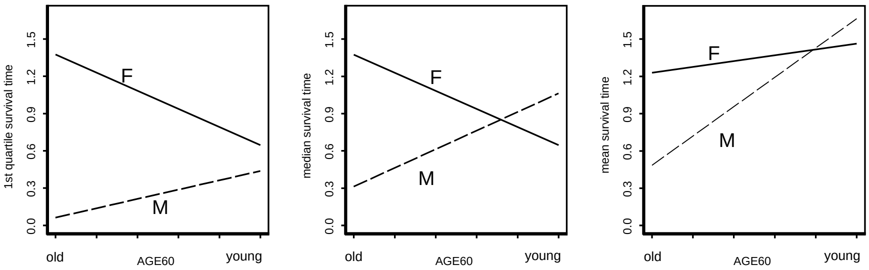

Figure 5.3: *Interaction between* SEX *and* AGE60 *adjusted for* KPS.PRE. and GROUP *via* `coxph` *and then evaluated at* GROUP = 1 *and* KPS.PRE. = 80.

p. 139: Exercise 5.2: "GROUP=1 and KPS.PRE.=70." Remove the entire **Tip:**

p. 143: Item 5: provides `deviance` residuals, `dfbeta`, and `dfbetas`.

p. 145: L-2: "two sample" should be "two-sample".

p. 146: L3 to L4: Delete "Hence, the plots of the ratios are horizontal lines."

p. 147:

Replace the first paragraph under **Graphical checks for overall model adequacy** with the following:

We see that $\log(t_p)$ is not only linear in z_p, but also in each $x^{(j)}$, $j = 1, \ldots, m$. Further, the above linear model says $(Y - \beta_0^* - \underline{x}'\underline{\beta}^*)/\sigma = Z$. Define the ith residual e_i to be

$$e_i = \frac{y_i - \widehat{y}_i}{\widehat{\sigma}} \, ,$$

where $\widehat{y}_i = \widehat{\beta}_0^* + \underline{x}'\underline{\widehat{\beta}}^*$ is the ith estimated linear predictor. Under the Weibull model, the set of uncensored residuals should behave roughly like a set of iid standard extreme value variates. Let $e_1, e_2, \ldots, e_r$, $r \leq n$, represent the ordered uncensored residuals. We draw a Q-Q plot (page 63) of the points (z_i, e_i), $i = 1, \ldots, r \leq n$. In the recipe given on page 63, replace the sample quantile y_i with e_i and proceed to obtain the corresponding parametric quantile z_i. If the model under study (here it is the Weibull) is appropriate, the points (z_i, e_i) should lie very close to the 45^o-line through the origin. Figure 6.4 displays the Q-Q plot. Lastly, draw m scatter plots of the points $(x_i^{(j)}, y_i)$, $i = 1, \ldots, r \leq n$ and $j = 1, \ldots, m$. Each plot should display a straight line pattern. If not, perhaps transforming those $x_i^{(j)}$'s could improve the fit. See Figure 6.5.

The function `qq.reg.resid.s` (for R users: `qq.reg.resid.r`) draws a Q-Q plot of the e_i residuals. It has six arguments. They are:

```
data = data.frame
time = survival time variable name in data.frame
status = name of status variable in data.frame
fit = a survReg object
quantile = "qweibull" or "qnorm" or "qlogis"
xlab = "type your label" E.g., "standard extreme value quantiles"
```

S code for Q-Q plot of (z_i, e_i) after fitting the Motorette data to a Weibull regression model:

```
> fit.weib <- survReg(Surv(time,status) ~ x,dist="weibull",
              data=motorette)
> qq.reg.resid.s(motorette,motorette$time,motorette$status,fit.weib,
              "qweibull","standard extreme value quantiles")
              # Produces Figure 6.4
```

p. 148:

L3 from expression (6.6): Replace the sentence "A saturated model . . . perfectly." with: A saturated model in the regression setting without censoring is one with n parameters that fit the n observations perfectly. But in the presence of censored data, one needs to be careful. In view of (1.13), the factors of the likelihood corresponding to censored observations entail maximizing the survival probability.

L4 from expression (6.6): Replace the sentence "This also entails . . . **no** constraints." with: This entails that for uncensored observations we obtain the MLE's with **no** constraints; whereas for censored observations, maximizing a survival probability imposes a constraint on the θ_i's fit to these censored y_i's.

p. 150:

Right after (6.12): Remove "where the weight ... uncensored points."

In addition, remove $w_i\times$ from D_i in expression (6.12), and insert a period right after D_i.

p. 152:

L4: Replace (the reduced one) with R.

L7: Replace reduced model) with R) = SSE(R) − SSE(F)

L8: $\log(T_i)$

L9: simplifies to

$$n \log\left(\frac{\text{MSE(R)}}{\text{MSE(F)}}\right) + (P_F - P_R),$$

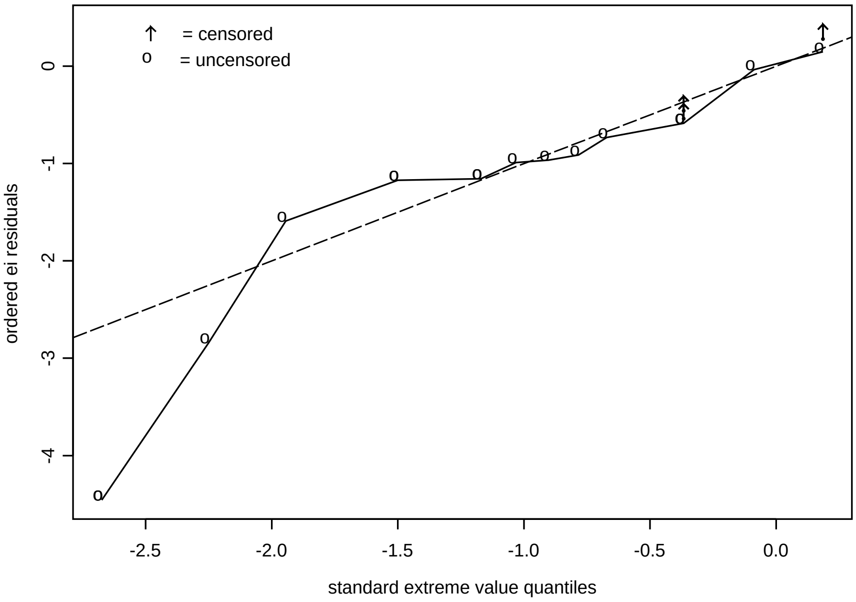

Figure 6.4: *Q-Q plot for the e_i residuals. Dashed line is the 45^o-line.*

where P_F and P_R are the numbers of parameters in the full and reduced models, respectively. The argument of the log function can be easily expressed as a function of the classic F test statistic to test a reduced model against the full model. Thus, $\mathcal{PD}$ simplifies to an increasing function of the classic F statistic, which has in its numerator the extra sum of squares SSE(R) $-$ SSE(F). WHY!

p. 155: Figure 6.4 is wrong. Replace it with Figure 6.4 printed in this errata sheet.

p. 156:

Figure 6.10, upper left plot: Replace x-axis caption “riskscore” with “risk scores”.

Figure 6.10, upper right plot: Replace x-axis caption “Quantiles of Standard Normal” with “normal scores”.

Figure 6.10 caption: replace “*quantiles*” with “*scores*”.

Third bullet: Replace the entire paragraph by the following: Figure 6.5 displays a straight line. Figure 6.7 shows no influential points. Both Figure 6.6 and

Figure 6.10 (deviance residuals vs. risk scores) display a random scatter about zero except for a possible outlier whose deviance residual value is -2.634, which, incidentally, represents the two extreme cases detected by the deviance residual vs. index plot. These two cases correspond to the possible outlier revealed in the Q-Q plot displayed in Figure 6.4.

Insert in between the third and fourth bullets this new bullet:

• The plot of deviance residuals against their normal scores in Figure 6.10 suggests one potential outlier. But this is somewhat misleading. The three upper right points correspond to cases with the same deviance residual value of 1.626052, but with different normal scores. This occurs because the S function `qqnorm` assigns these residuals their distinct ranks $k = 28, 29$, and 30. Hence, their normal scores $(Z((k - .375)/(n + .25)))$ are 1.361, 1.61, and 2.04, respectively. However, if we follow the convention of assigning the average rank to tied observations, then each of these three tied deviance residuals now has the normal score value of 1.61. In this case, the three points are now the single point in the middle and there are no apparent outliers in this plot.

p. 163: L4 from Remarks item 1: the log-partial likelihood

p. 166:

L2 from subsection 6.3.7: cns2.coxint6, page 128.

L-2 from Figure 6.12: “plots” should be “plot”

L-1 from Figure 6.12: Replace “follow” with “follows”

p. 167: L-6 from Figure 6.15: `#` Schoenfeld resid for KPS.PRE.

p. 169: Insert “In Figure 6.14” between the second bullet and “the subjects . . .”

p. 171: The x-axis on the plot in Figure 6.14: Ordered survival time

p. 173:

L8 from the caption of the Figure 6.17: profile likelihood function should be profile log-likelihood function.

S commands on the bottom: Replace them with the following:

```
> cns2.temp <- move.col(cns2,1,cns2,9) # Moves KPS.PRE. in col. 9
                                       # to col. 1
> cns2.temp$K <- rbinom(nrow(cns2),1,.5) # Creates a place holder
                                       # for K
> temp <- coxph(Surv(B3TODEATH,STATUS) ~ GROUP+SEX+AGE60+
            SEX:AGE60+K, data=cns2.temp) # provides the necessary
```

p. 174:

L2: Replace `data=cns2` with `data=cns2.temp`

The first two S commands on the bottom: Replace with the following:

```
> cns2.temp$K <- as.integer(cns2.temp$KPS.PRE.>=90) # creates the
                                     # indicator variable K
> cns2cutpt1 <- coxph(Surv(B3TODEATH,STATUS) ~ GROUP+SEX+
                         AGE60+SEX:AGE60+K,data=cns2.temp)
```

p. 175: The last line of the S commands: Delete `> detach()`

p. 176: First S command: Replace cns2 with cns2.temp. That is, type:

```
> temp1 <- bootstrap(cns2.temp,cutpt.coxph(object=cns2cutpt1,
              data=cns2.temp,q=seq(.05,.95,.05))$cutpt,B=1000)
```

p. 177:

L2 from Figure 6.20 caption: ... main effects model cns2.cox3, page 130.

In Table 6.2, p-value for KPS.PRE. should be .00056

p. 179: Exercise 6.2: Remove "(Continued from Exercise 6.1)". Insert "in your chosen model in Exercise 5.1(c)" right after `alter`. Same line: Exercise 5.1(a) should be Exercise 5.1(c). Insert "Was the martingale residual plot for `alter` helpful in finding a cut point?"

p. 193:

L1 and L12: Andersen-Gill

L-11: exposure variable, status variable, and the time variable into columns 2, 3, and 4, respectively.

p. 197:

L-4: $t < T_M < t + \Delta t$ should be $t \leq T_M < t + \Delta t$

L-5: $T_M > t$ should be $T_M \geq t$

p. 199:

L5: Replace KM(t)$_{12}$ by KM$_{12}(t)$

L10: n^*_{j-1}

p. 200:

L11: Replace m by $m-1$.

L14: In CI(t), replace the index z_j only under symbol Σ by z_{j+1}

L15: Replace $j = 0\ldots, m \leq n-1$ by $j = 0, \ldots, m-1 \leq n-1$

L24: In $J_{\text{KM}}(z_{j+1})$, replace rj by r_j

L26: In 1 - KM(t), replace the index z_j only under symbol Σ by z_{j+1}.

p. 201: In Table 7.5, replace $J_{\text{CI}}(z_{j+1})$ and $J_{\text{KM}}(z_{j+1})$ by $J_{\text{CI}}(z_j)$ and $J_{\text{KM}}(z_j)$, respectively.

p. 203:

L4 from the end of item 4: a survivor function of C

L-7: In the displayed expression $L = \cdots$, the term $S_f(x)$ appears in three denominators. Replace each with $S_f(x_i)$.

p. 204:

Expression (7.8):

$$P(x, y, \delta) = \frac{\big(f(y)\big)^{\delta}\big(S_f(y)\big)^{1-\delta}}{S_f(x)} = \frac{\big(h(y)\big)^{\delta} S_f(y)}{S_f(x)},$$

L1 from expression (7.8): Insert the following after T.: The missing factor $k(x)\big(S_g(y)\big)^{\delta}\big(g(y)\big)^{1-\delta}$ is considered to be a constant as it plays no role in the maximization process of the conditional likelihood L. This results from the assumption that the distributions of X and C do not involve any parameters of interest.

L-5: from an exponential distribution

p. 207: L3: Insert "survival data for female psychiatric patients" right after presents.

p. 221: Figure 8.5: The second and third lines are 2x + y = 5.3 and 2x + y = 5.8.

p. 223: L1: "next section" should be "Section 8.5"

p. 225: L2 from expression (8.13): Remove the braces from $\{\underline{x}_i\}$.

p. 227: L3 from Section 8.5.1: change “empirical df” to “empirical survival function”.

p. 233: Insert at the beginning of the R commands at bottom of page

```
> library("crq",lib.doc="/your directory/library) # loads libraries
                               # crq, survival, and modreg
```

L-1: `> ex <- crq(Surv(log(y),s)~ x, mw=1)`

p. 235: R commands in the middle: Remove `> attach(cns2)` and replace the next line with the following:

```
> cns2.crq <- crq(Surv(log(B3TODEATH),STATUS) ~ KPS.PRE. + GROUP
                          + SEX, data=cns2, mw=0)
```

p. 236:

The first R commands:

```
> Kd <- cns2$KPS.PRE. + .0001*rnorm(58)
> Gd <- cns2$GROUP + .0001*rnorm(58)
> Sd <- cns2$SEX + .0001*rnorm(58)
> dith <- crq(Surv(log(B3TODEATH),STATUS) ~ Kd+Gd+Sd,data=cns2)
> cns2.b <- crq.boot(dith,400,.1*(1:8))
```

The R commands after expression (8.20):

```
> cns2.cox <- coxph(Surv(B3TODEATH,STATUS) ~ KPS.PRE.+GROUP
                       +SEX, data=cns2)
> pltbeta(cns2.crq,cns2.b,2,cns2.cox,"CNS2",acc=TRUE)
> pltbeta(cns2.crq,cns2.b,2,cns2.cox,"CNS2",acc=TRUE)
> pltbeta(cns2.crq,cns2.b,3,cns2.cox,"CNS2",acc=TRUE)
```

The “CNS2” in the calling sequence is just for the title.

p. 238: Stephen Portnoy has put up a new version of the censored regression quantile programs on his website `www.stat.uiuc.edu/~portnoy/steve.html`. It does some cosmetic improvements and corrects an “error” that he discovered in the Bootstrap method he was using. It is something done by many (most) people who use the bootstrap, but it is subtly wrong. It is sometimes called the naive bootstrap: If you have an estimator, $\hat{\theta}$, of θ; and take B bootstrap samples getting B $\theta^\star$'s, then you might be tempted to use as a 95% C.I. the interval, $(\theta^\star_{.025}, \theta^\star_{.975})$. Actually, this is “first order” correct. But it is wrong in the sense that it won't pick up the higher order corrections the bootstrap is supposed to get. The reason is that the distribution of $(\theta^\star - \hat{\theta})$ is what approximates the distribution of $(\hat{\theta} - \theta)$ (to higher order). Thus, we need to solve the inequalities

$$\theta^\star_{.025} - \hat{\theta} < \hat{\theta} - \theta < \theta^\star_{.975} - \hat{\theta}$$

to get the confidence interval:

$$(\hat{\theta} - (\theta^{\star}_{.975} - \hat{\theta}), \hat{\theta} - (\theta^{\star}_{.025} - \hat{\theta}))$$

which simplifies to

$$(2 \times \hat{\theta} - \theta^{\star}_{.975}, 2 \times \hat{\theta} - \theta^{\star}_{.025}).$$

In the earlier version of the software he had used the incorrect "naive" bootstrap based on the IQR. This is now corrected. It does mean that anyone using the new version will get slightly different confidence intervals from those in the book (any asymmetries will be reversed), but the p-values should be the same.

```
> sumpr.crq(ci.boot(cns2.crq,cns2.b,c(3,5,7)))
```

p. 242 - 244: Here the data set "addicts" is created; the difference being, the variable Days.survival in Table 7.2, page 188, is renamed by Time.

```
> attach(addicts)
> add.crq <- crq(Surv(log(Time),Status) ~ Clinic + Prison + Dose)
> cd <- Clinic + .0001*rnorm(238)
> pd <- Prison + .0001*rnorm(238)
> dd <- Dose + .0001*rnorm(238)
> dith <- crq(Surv(log(Time),Status) ~ cd + pd + dd)
> add.b <- crq.boot(dith, 400, .05*(1:13))
> sumpr.crq(ci.boot(add.crq,add.b,c(5,8,10,12)))
> add.cox4 <- coxph(Surv(Time,Status) ~ Clinic + Prison
                     + log(Dose))
> add.crq4 <- crq(Surv(log(Time),Status) ~ Clinic + Prison
                   + log(Dose), mw=100)
> add.b4 <- crq.boot(add.crq4,400,.05*(1:13),mw=100)
> summary(coxph(Surv(Time,Status) ~ Clinic + Prison + log(Dose)))
> print(cox.zph(add.cox4))
> sumpr.crq(ci.boot(add.crq4,add.b4,c(5,8,10,12)))
```

Note: Occasionally `attach` command is unstable. To avoid this, delete it and insert `data=addicts` where necessary.

p. 249: Switch the order of the Kaplan and Kalbfleisch references.

p. 251: Andersen-Gill formulation