C.6 Modern Survival Problem sheet 6: Censoring and truncation, frailty and repeated events

(1) A sample of patients taking a new blood pressure medication is asked whether they have experienced any vertigo since they started taking it; and if so, when the symptoms were first noticed. Some have not experienced symptoms yet, some report an exact time (in weeks after starting treatment), and some can only say they know it was before a certain time.

Which observations are left-censored? Right-censored? Estimate the survival function (that is, probability of remaining symptom-free for x weeks)

The first column are the right-censored observations. The second column are the left-censored observations.

(a) Ignoring the left-censored observations;

weeks	# at risk	p_x	$\hat{S}(x)$
1	307	0.980	0.980
2	256	0.957	0.938
3	223	0.955	0.896
4	190	0.884	0.792
5	149	0.752	0.596
6	100	0.670	0.399
7	57	0.719	0.287
8	38	0.658	0.189
9	20	0.600	0.113
10	9	0.000	0.000

(b) Ignoring the right-censored observations;

We reverse time from 11 weeks. Letting T be the time of first vertigo, and $\tau_i = 11 - T_i$. we compute a Kaplan-Meier survival estimator for τ.

weeks	# at risk	p_x	$\hat{S}_\tau(x)$
1	213	0.958	0.958
2	189	0.958	0.917
3	172	0.924	0.848
4	155	0.897	0.760
5	130	0.746	0.567
6	91	0.593	0.337
7	52	0.577	0.194
8	27	0.630	0.122
9	17	0.353	0.043
10	6	0.000	0.000

Now, $\hat{S}_\tau(x)$ is an estimator for

$$\mathbb{P}\{11 - T > x\} = \mathbb{P}\{T < 11 - x\} = 1 - \mathbb{P}\{T \geq 11 - x\} = 1 - S_T(11 - x+).$$

Thus we can estimate

$$S_T(y) \approx 1 - \hat{S}_\tau(11 - y-) = 1 - \hat{S}_\tau(10 - y);$$

that is, the estimate of survival for τ just before time $11 - y$, which in this case is the same as the survival estimate at time $10 - y$, since there is no change between integer times. This yields

weeks	$\hat{S}_T(x)$
1	0.957
2	0.878
3	0.806
4	0.663
5	0.433
6	0.240
7	0.152
8	0.083
9	0.042
10	0.000

Note that there is a certain amount of ambiguity here in the way we have dealt with the discreteness of the censoring and event times.

(c) Taking all observations into account.

We apply Turnbull's algorithm for doubly censored data, beginning with the solution from part (a), calling that $\hat{S}_0$. Because the data come at discrete times, the grid of times will just be $1, 2, 3, 4, 5, 6, 7, 8, 9, 10$.

We compute

$$p_{j\ell}^{(0)}(0) = \frac{\hat{S}_0(t_{\ell-1}) - \hat{S}_0(t_\ell)}{1 - \hat{S}_0(t_j)}$$

$$= \begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.32 & 0.68 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.19 & 0.41 & 0.41 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.09 & 0.20 & 0.20 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.05 & 0.10 & 0.10 & 0.26 & 0.49 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.03 & 0.07 & 0.07 & 0.17 & 0.33 & 0.33 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.03 & 0.06 & 0.06 & 0.15 & 0.28 & 0.28 & 0.16 & 0.00 & 0.00 & 0.00 \\ 0.02 & 0.05 & 0.05 & 0.13 & 0.24 & 0.24 & 0.14 & 0.12 & 0.00 & 0.00 \\ 0.02 & 0.05 & 0.05 & 0.12 & 0.22 & 0.22 & 0.13 & 0.11 & 0.09 & 0.00 \\ 0.02 & 0.04 & 0.04 & 0.10 & 0.20 & 0.20 & 0.11 & 0.10 & 0.08 & 0.11 \end{pmatrix}.$$

Note that we have treated left-censoring at time t as meaning $T \leq t$. This may seem inappropriate: Perhaps someone who at week 3 cannot recall when symptoms began should be assumed to have started them in weeks 1 or 2. On the other hand, it is plausible that someone would have symptoms beginning during week 3, but would report at the end of week 3 that she can't remember which week they started in. This is part of the more general problem, that our modelling assumption of non-informative left censoring is probably not very appropriate to this story.

Now, we reassign the left-censored observations according to this distribution, obtaining the following numbers of estimated events:

weeks	number of events
1	7.40
2	14.00
3	13.00
4	29.50
5	48.30
6	43.40
7	20.80
8	16.00
9	9.90
10	10.70

Computing the Kaplan–Meier estimator again, with these new numbers of events, we get

weeks	# at risk	p_x	$\hat{S}_\tau(x)$
1	355	0.979	0.979
2	302	0.954	0.934
3	266	0.951	0.888
4	230	0.872	0.774
5	182	0.734	0.569
6	121	0.644	0.366
7	68	0.696	0.255
8	44	0.642	0.164
9	23	0.581	0.095
10	10	0.000	0.000

This is our second estimator, $\hat{S}_1$. It is slightly different from $\hat{S}_0$. We can iterate the procedure, carrying out exactly the same calculation with $\hat{S}_1$ in place of $\hat{S}_2$. The new estimated numbers of events are

weeks	number of events
1	7.41
2	14.04
3	13.03
4	29.48
5	48.34
6	43.36
7	20.78
8	15.95
9	9.90
10	10.70

The redistribution is minuscule, so it is probably not worth continuing with another iteration of the survival estimation.

(2) In order to control the spread of a virus in a wild population, researchers spread food items laced with a vaccine. Once a week they capture a small number of animals and test whether they have developed an immune response

week	1	2	3	4	5	6	7	8	9	10
number sampled	5	4	7	3	4	6	3	8	5	4
number immune	0	1	2	0	2	1	2	4	4	3

Estimate the probability of being immune at week t

(a) using an exponential model;
sec:currentstatusparametric
Using the results in section 15.2.1 we have the log likelihood

$$\ell(\lambda) = \sum_{c=1}^{10} k_c \log\left(1 - e^{-\lambda c}\right) - \lambda \sum_{c=1}^{10} (n_c - k_c)c,$$

where n_c is the number of animals sampled at week c, and k_c the number found to be immune. We can find the maximum numerically:

```
n=c(5,4,7,3,4,6,3,8,5,4)
k=c(0,1,2,0,2,1,2,4,4,3)


loglik=function(lambda){
  c=1:10
  -sum(k*log(1-exp(-lambda*c))-lambda*c*(n-k))
}
# Note: This is negative log likelihood because optimize finds minima

optimize(loglik,c(0,2))
$minimum
[1] 0.097283

$objective
[1] 27.88519
```

So $\hat{\lambda} = 0.097$.

(b) using a Weibull model;
The Weibull log likelihood with cumulative hazard parametrised as $\Lambda(t) = (\lambda t)^r$ is

$$\ell(\lambda, r) = \sum_{c=1}^{10} k_c \log\left(1 - e^{-(\lambda c)^r}\right) - \lambda^r \sum_{c=1}^{10} (n_c - k_c)c^r.$$

We use the `nlm` function to minimise a function of two variables. (We need to give a starting point, for which we take the exponential solution that we found in the previous example).

```
loglik=function(lambda,r){
  c=1:10
  -sum(k*log(1-exp(-(lambda*c)^r))-(lambda*c)^r*(n-k))
}
> nlm(function(x) loglik(x[1],x[2]),c(.1,1))
$minimum
[1] 27.52188
```

```
$estimate
[1] 0.1135262 1.4416975

$gradient
[1] 9.006129e-09 3.998258e-08

$code
[1] 1

$iterations
[1] 11
```

Thus the MLE for the Weibull distribution has parameters $(\hat{\lambda}, \hat{r}) = (0.11, 1.44)$. We note that the log likelihood has been increased only from -27.8 to -27.5, so by the likelihood ratio test we would not take the Weibull distribution as an improvement.

(c) using the nonparametric MLE.

We apply the Pool Adjacent Violators Algorithm. We start by calculating the fraction "surviving" at each census time

week	1	2	3	4	5	6	7	8	9	10
number sampled	5	4	7	3	4	6	3	8	5	4
fraction not yet immune	1.0	0.75	0.71	1.0	0.5	0.83	0.33	0.50	0.20	0.25

We see that weeks $(3, 4)$, $(5, 6)$, $(7, 8)$, and $(9, 10)$ are all increasing. So we pool these observations:

week	1	2	3 4	5 6	7 8	9 10
number sampled	5	4	10	10	11	9
fraction not yet immune	1.0	0.75	0.8	0.7	0.45	0.22

There remains one increasing sequence, so we pool $(2, 3, 4)$:

week	1	2 3 4	5 6	7 8	9 10
number sampled	5	14	10	11	9
fraction not yet immune	1.0	0.79	0.7	0.45	0.22

(3) A population has multiplicative frailty, so that the mortality rate for individual i is $B_i\alpha(x)$ at age x, where the B_i are i.i.d. positive random variables, where $\lim_{x\to\infty} \alpha(x) = \infty$.

(a) Show that the population mortality goes to ∞ as $x \to \infty$ if the distribution of B_i is bounded away from 0.

Suppose $B_i \geq b > 0$ with probability 1. Since $e^{-B_i\theta} > 0$,

$$\frac{-\mathcal{L}'(\theta)}{\mathcal{L}(\theta)} = \frac{\mathbb{E}\left[B_i e^{-B_i\theta}\right]}{\mathbb{E}\left[e^{-B_i\theta}\right]} \geq \frac{\mathbb{E}\left[be^{-B_i\theta}\right]}{\mathbb{E}\left[e^{-B_i\theta}\right]} \geq b.$$

By equation (16.2), it follows that the population mortality rate is bounded below by $b\alpha(x)$, which goes to ∞.

(b) Show that the population mortality converges to a finite constant as $x \to \infty$ if the distribution of B_i has nonzero density at 0 and the hazard rate does not grow too quickly as $x \to \infty$. Give a formal condition for what "too quickly" would be.
Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be the density of B_i, with $f(0) > 0$. For large θ, $e^{-\theta x} f(x)$ is almost the same as $e^{-\theta x} f(0)$ (since $e^{-\theta x} f(x)$ is nearly 0 except for $x \approx 0$), so

$$\frac{-\mathcal{L}'(\theta)}{\mathcal{L}(\theta)} = \frac{\int_0^\infty x e^{-\theta x} f(x) dx}{\int_0^\infty e^{-\theta x} f(x) dx} \approx \frac{f(0)}{\theta}.$$

Thus, by (16.2) the population mortality for large x is

$$\mu(x) \approx f(0) \frac{\alpha(x)}{A(x)}.$$

This will be bounded, unless $\alpha(x)$ grows extremely fast. The condition that needs to be satisfied is that

$$\lim_{x\to\infty} \int_0^x \frac{\alpha(y)}{\alpha(x)} dy > 0.$$

This will certainly be true if α is a Gompertz hazard (so grows exponentially with x).

(c) Suppose now that the baseline hazard is Gompertz, i.e., $\alpha(x) = e^{\theta x}$.

i. If the B_i have Gamma distribution with parameters (r, λ) — λ is the rate parameter — compute the population mortality rate $\mu(t)$ at age t.
The Laplace transform is $\mathcal{L}(c) = (1 + c/\lambda)^{-r}$. Thus the population mortality is

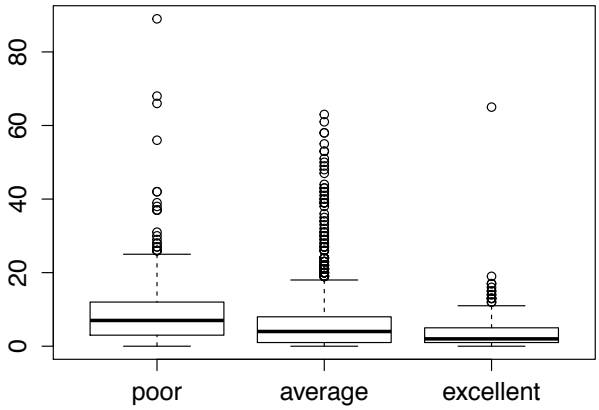
$$\mu(x) = \frac{re^{\theta x}}{\lambda}(1 + (e^{\theta x} - 1)/\theta\lambda)^{-1} = \frac{\theta r}{(\theta\lambda - 1)e^{-\theta x} + 1}.$$

ii. What is the hazard ratio between a subpopulation whose frailty has Gamma distribution with parameters (r, λ) and one with parameters (r', λ)?
From the above formula it will be r/r'.

(4) The paper [ZKJ07] includes a dataset, available to download from the Journal of Statistical Science, on the healthcare demand of 4406 patients in the public old-age health insurance scheme Medicare in the US. When you load this file in, the data will be in a data-frame `DebTrivedi`.

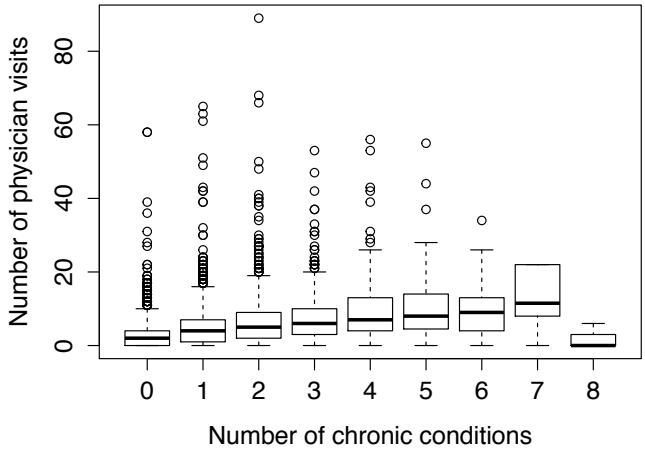
(a) The number of physician office visits is enumerated in the variable `ofp`, while `numchron` gives the number of chronic conditions, and `health` gives self-reported health (poor, average, excellent). Do one or more exploratory plots to illustrate the distributions of these variables, and their relationship.

```
attach(DebTrivedi)
# box plot of visits by health status
boxplot(ofp~health)
# box plot of visits by number of conditions
boxplot(ofp~factor(numchron),xlab='Number of chronic conditions',ylab='Number
    of physician visits')
# histogram (bar plot) of number of conditions, stratified by health
plot(-1,-1,xlim=c(-.5,8.5),ylim=c(0,1),xlab='Number of chronic conditions',
   ylab='fraction')
c=-1
for(L in levels(health)){
  h=hist(numchron[health==L],breaks=0:9,plot=FALSE)
  rect((0:8)+c/3-1/6,0,(0:8)+c/3+1/6,h$density,col=c+3)
  c=c+1
}
legend(4,.6,c('poor health','average health','excellent health'),col=2:4,lwd
   =3)
```



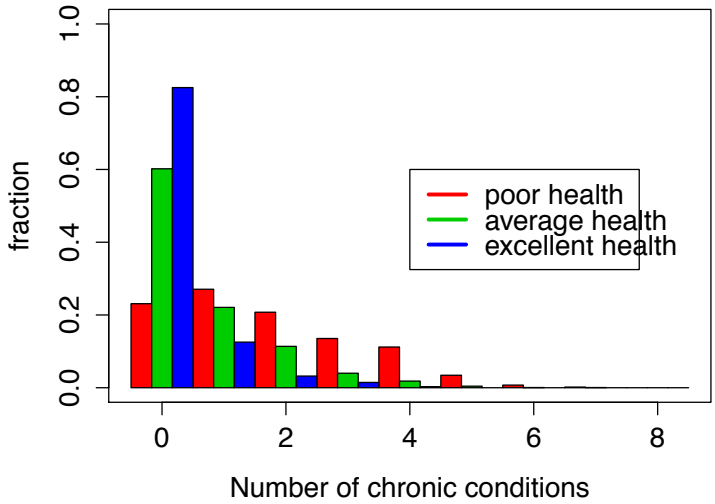
F:boxvisitshealth

Figure C.6: Box plot of visits by health status.



F:boxvisitsnumchron

Figure C.7: Box plot of visits by number of conditions.



F:histnumcon

Figure C.8: Histograms of number of chronic conditions, stratified by health status.

(b) Fit a Poisson regression model to predict the number of office visits as a function of `health`, `numchron`, `gender`, `school` (number of years of schooling), and `privins` (indicator of whether the patient has private insurance). Interpret the result.

```
> preg=glm(ofp~health+numchron+gender+school+privins,family=poisson)
> summary(preg)

Call:
glm(formula = ofp ~ health + numchron + gender + school + privins,
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.2816  -2.0370  -0.7143   0.7301  16.2655

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.034542   0.023857  43.364   <2e-16 ***
healthpoor      0.318205   0.017479  18.205   <2e-16 ***
healthexcellent -0.379045   0.030291 -12.514   <2e-16 ***
numchron        0.168793   0.004471  37.755   <2e-16 ***
gendermale      -0.108014   0.012943  -8.346   <2e-16 ***
school          0.025754   0.001843  13.972   <2e-16 ***
privinsyes      0.216007   0.016872  12.803   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26943  on 4405  degrees of freedom
Residual deviance: 23808  on 4399  degrees of freedom
AIC: 36597

Number of Fisher Scoring iterations: 5
```

All of these effects seem to be significant. Unsurprisingly, excellent health is associated with a reduction in the rate of physician visits, and poor health with an increase. Male patients have slightly fewer (by a factor of $e^{-0.108} = 0.898$). More schooling and private insurance are both associated with an increase in the number of office visits.

(c) Explain why you might want to fit a negative binomial model instead. Do the fit, and interpret the result.

We would expect individuals to have differing propensities to go to see a physician, separate from the factors included in the model. And the tail of `ofp` seems much too long (`ofp` is **overdispersed** relative to Poisson) to be explained by any Poisson distribution.

```
> preg2=glm.nb(ofp~health+numchron+gender+school+privins,link=log)
> summary(preg2)

Call:
glm.nb(formula = ofp ~ health + numchron + gender + school +
    privins, link = log, init.theta = 1.164195333)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6730  -1.0062  -0.3002   0.2859   5.6124

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.940307   0.055296  17.005  < 2e-16 ***
healthpoor      0.367665   0.048733   7.544 4.54e-14 ***
healthexcellent -0.373647   0.061669  -6.059 1.37e-09 ***
numchron        0.195760   0.012067  16.223  < 2e-16 ***
gendermale      -0.115130   0.031609  -3.642  0.00027 ***
school          0.027179   0.004451   6.106 1.02e-09 ***
privinsyes      0.250154   0.040008   6.253 4.04e-10 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for Negative Binomial(1.1642) family taken to be 1)

    Null deviance: 5607.2  on 4405  degrees of freedom
Residual deviance: 5039.2  on 4399  degrees of freedom
AIC: 24470

Number of Fisher Scoring iterations: 1


              Theta:  1.1642
          Std. Err.:  0.0320

 2 x log-likelihood:  -24453.9070
```

We see a huge reduction in AIC, indicating a superior fit. The deviance residuals are much more controlled. On the other hand, the parameter estimates remain fairly similar.