Problem 1 *Hospital length of stay of coronavirus (COVID-19) patients*

A group of Chinese medical researchers analyzed data for length of hospital stay for confirmed COVID-19 patients at hospitals in the Sichuan Province. The aim of the study was to obtain knowledge about the new virus that would be important for planning and allocation of medical resources in the COVID-19 pandemic.

The study included 538 patients who were admitted in hospitals after January 16, 2020. 351 out of these (65%) recovered and were discharged before the end of the study, April 4. Only 3 patients died in hospital before April 4.

The data used in this exercise are *simulated* based on the reported results from the study (the full data set was not published in the report).

The data consist of the observed time, `Time` (in days); censoring status C (= 0 or 1); and six binary covariates $\mathtt{x}_1, \ldots, \mathtt{x}_6$, with values 0 and 1 as defined by Table 1.

For patients that were discharged at or before April 4, `Time` is the true length of hospital stay. These patients are given censoring status $C = 1$. For patients that were alive and still in hospital on April 4, `Time` is the observed length of stay in hospital. These patients are considered as censored and given censoring status $C = 0$. For patients that died in hospital, `Time` is the number of days until death, and censoring status is again $C = 0$.

In Table 1, 'time from onset' means time from onset of COVID-19 to admission at the hospital; 'hospital grade' distinguishes between admission to *provincial* and *non-provincial* hospitals; 'density of health workers' means number of health workers per 1000 inhabitants; 'clinical grade' is degree of illness.

i	x_i	0	1
1	age (years)	< 45	≥ 45
2	gender	male	female
3	time from onset	< 5	≥ 5
4	hospital grade	non-provincial	provincial
5	density of health workers	< 5.5	≥ 5.5
6	clinical grade	mild	severe

Table 1: The binary covariates

The data for a randomly selected subset of 16 of the 538 patients are displayed in Figure 1 on the next page.

Row	x1	x2	x3	x4	x5	x6	Time	C
1	0	1	0	0	1	1	3	0
2	1	1	1	0	1	0	4	0
3	0	1	1	0	1	0	6	1
4	1	1	0	0	1	0	8	0
5	0	0	1	0	1	0	10	1
6	0	0	0	0	0	0	12	1
7	0	0	0	1	0	1	12	0
8	1	0	0	0	0	0	13	0
9	0	1	0	0	0	0	13	1
10	0	0	1	0	0	0	14	0
11	1	1	0	0	0	0	18	1
12	0	0	1	0	0	1	18	1
13	1	1	0	0	1	0	19	1
14	0	1	0	0	1	0	21	1
15	1	1	0	0	0	0	22	1
16	1	0	1	0	0	0	26	0

Figure 1: The data for 16 patients

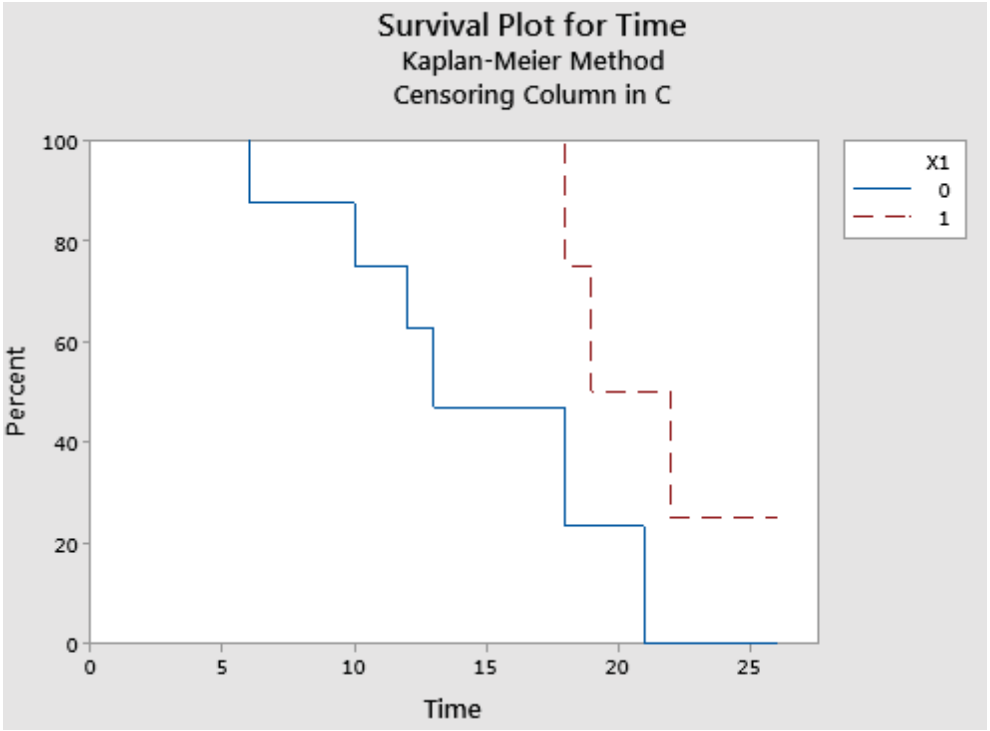

Figure 2: Kaplan-Meier plots (Minitab) for length of hospital stay for the 16 patients in Figure 1. *Solid line:* Age < 45. *Dashed line:* Age ≥ 45

a) Figure 2 shows Kaplan-Meier plots for the length of stay for the two age groups < 45 and ≥ 45 (i.e., $x_1 = 0$ and 1), using the data for 16 patients in Figure 1. Here, covariates $x_2, \ldots, x_6$ are ignored.

Use the data in Figure 1 to do the calculations leading to the two Kaplan-Meier plots.

How would you estimate the mean lengths of stay for the two age-groups based on the plots? You need not do the full calculations. (*Answer:* For age < 45: 14.67; for age ≥ 45: 21.25).

What conclusion can you draw from this limited study regarding the influence of age on length of hospital stay?

The following is part of the output of a *Weibull*-regression in Minitab, using data from all the 538 patients, and all the covariates $x_1, \ldots, x_6$.

```
Distribution: Weibull

Regression Table
```

Predictor	Coef	Standard Error	Z	P	95,0% Normal CI Lower	Upper
Intercept	3,01557	0,0428301	70,41	0,000	2,93163	3,09952
X1	0,193824	0,0368677	5,26	0,000	0,121565	0,266083
X2	0,0322232	0,0366283	0,88	0,379	-0,0395670	0,104013
X3	-0,0698684	0,0374509	-1,87	0,062	-0,143271	0,0035341
X4	-0,0074103	0,0446739	-0,17	0,868	-0,0949695	0,0801488
X5	-0,102861	0,0367531	-2,80	0,005	-0,174895	-0,0308259
X6	0,214140	0,0522488	4,10	0,000	0,111734	0,316546
Shape	2,97223	0,118618			2,74860	3,21405

b) Give an interpretation of the estimated regression coefficients with respect to the effect of the corresponding covariate on the length of stay. What effects are significant? (Use significance level 5% when investigating significant covariates).

What is the estimated distribution of the length of stay for a patient with covariates $x_1, \ldots, x_6$?

Find an expression for the estimated median length of stay for this patient.

What is the relative increase in estimated median length of stay between a patient under 45 years and a patient over 45 years, when the other covariates are the same for the two?

(*Hint:* In general, the relative increase when going from a to b, where $0 < a \leq b$, is $(b - a)/a$.)

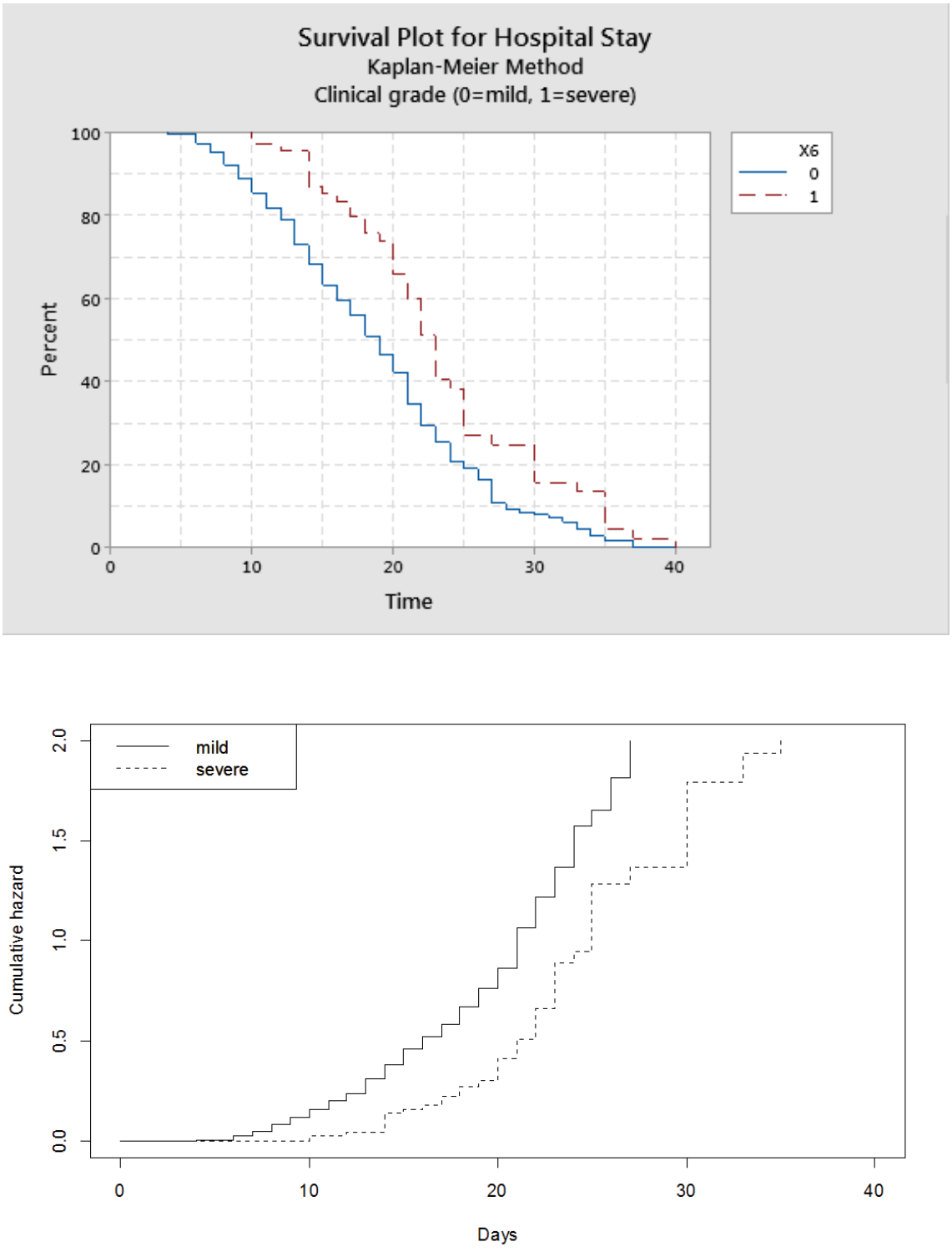

Figure 3: Kaplan-Meier plots using Minitab (upper panel) and Nelson-Aalen plots using R (lower panel) for the two groups of data corresponding to $x_6 = 0$ and $x_6 = 1$

The covariate for *clinical grade*, x_6 (0=mild, 1=severe), was of partiuclar interest in the study. In the rest of the exercise, we disregard the covariates $x_1, \ldots, x_5$ and instead group the 538 patients into two groups according to the value of x_6. This leads to a group of 463 patients with mild disease ($x_6 = 0$) and 75 patients with severe disease ($x_6 = 1$).

c) Figure 3 shows Kaplan-Meier plots and Nelson-Aalen plots for the two groups. Use the plots to do the following (give brief and rough answers; explanations are not needed):

1. Estimate the *median* length of stay in hospital for each of the two groups.
2. Estimate the *quartiles* for the length of stay for the two groups.
3. Make rough sketches of the *hazard functions* for the length of stay for each of the two groups.
4. Suppose that a certain patient with *severe* COVID-19 has been in hospital for 15 days. Give a rough estimate of the probability that this patient is discharged from hospital within the next day. Do the same for a patient with *mild* disease.

Below and on the next page are outputs of separate Weibull analyses in Minitab (*Parametric Distribution Analysis – Right Censoring*) for the two groups of patients defined by $x_6 = 0$ and $x_6 = 1$, respectively (using the 'By variable' option). Corresponding probability plots are shown in Figure 4 on the next page.

```
Distribution: Weibull

X6 = 0

Censoring Information  Count
Uncensored value         302
Right censored value     161

Parameter Estimates

                     Standard   95,0% Normal CI
Parameter  Estimate     Error    Lower    Upper
Shape       2,76367  0,119703  2,53874  3,00853
Scale       21,0908  0,442359  20,2414  21,9759

Log-Likelihood = -1077,416
```

```
X6 = 1

Censoring Information  Count
Uncensored value          49
Right censored value      26

Parameter Estimates

                     Standard   95,0% Normal CI
Parameter  Estimate     Error    Lower    Upper
Shape       3,62166  0,365623  2,97149  4,41408
Scale       26,0201   1,05255  24,0368  28,1670

Log-Likelihood = -169,687
```

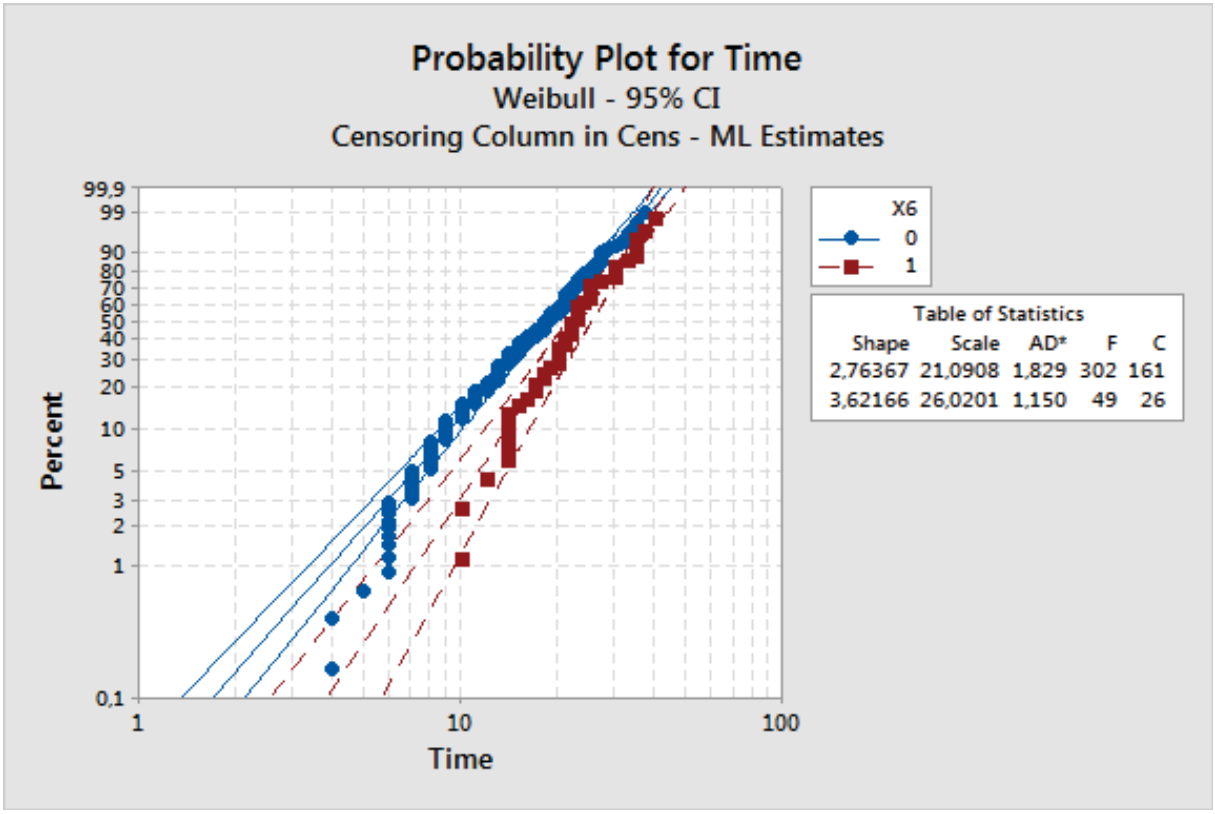

Figure 4: Probability plots for the two groups of data defined by $x_6 = 0$ and $x_6 = 1$

d) Looking at Figure 4, give a brief comment on the fit to Weibull distributions for each of the two groups defined by $x_6 = 0$ and $x_6 = 1$.

What aspect of the plotted points indicate that the shape parameters of the two Weibull distributions seem to differ?

For a formal comparison of the two shape parameters, you are asked to formulate an appropriate null hypothesis and a corresponding alternative hypothesis. Let, for example, α_0 and α_1 denote the shape parameters of the two groups corresponding to $x_6 = 0$ and $x_6 = 1$, respectively.

Perform the testing with significance level 5% by using information from the separate Weibull analyses for the two groups, as well as from the following output from a Weibull regression with the full data using covariate x_6 only:

```
Distribution: Weibull

Regression Table
```

Predictor	Coef	Standard Error	Z	P	95,0% Normal CI Lower	Upper
Intercept	3,05102	0,0202722	150,50	0,000	3,01129	3,09076
X6	0,192518	0,0538092	3,58	0,000	0,0870542	0,297982
Shape	2,86270	0,114186			2,64742	3,09548

```
Log-Likelihood = -1249,846
```

Problem 2 *Deaths of COVID-19 in Norway*

During the coronavirus pandemic, the Norwegian Institute of Public Health (FHI) records the number of deaths of COVID-19 in Norway each day. The table displayed in Figure 5 shows the number of dead for each of 63 days, starting March 11 (day 1) and ending May 12 (day 63). Figure 6 on the next page plots the cumulative number of deaths versus day number.

1	0	0	0	2	0	0	3	1	0
0	1	2	2	2	2	4	2	4	2
4	10	2	6	8	1	10	11	8	2
2	11	11	13	3	6	1	0	11	6
0	15	11	11	0	0	2	2	7	2
0	0	0	0	4	1	4	0	3	0
2	1	2							

Figure 5: Number of deaths by COVID-19 in Norway from March 11 (day 1) to May 12 (day 63). The table should be read row-wise.

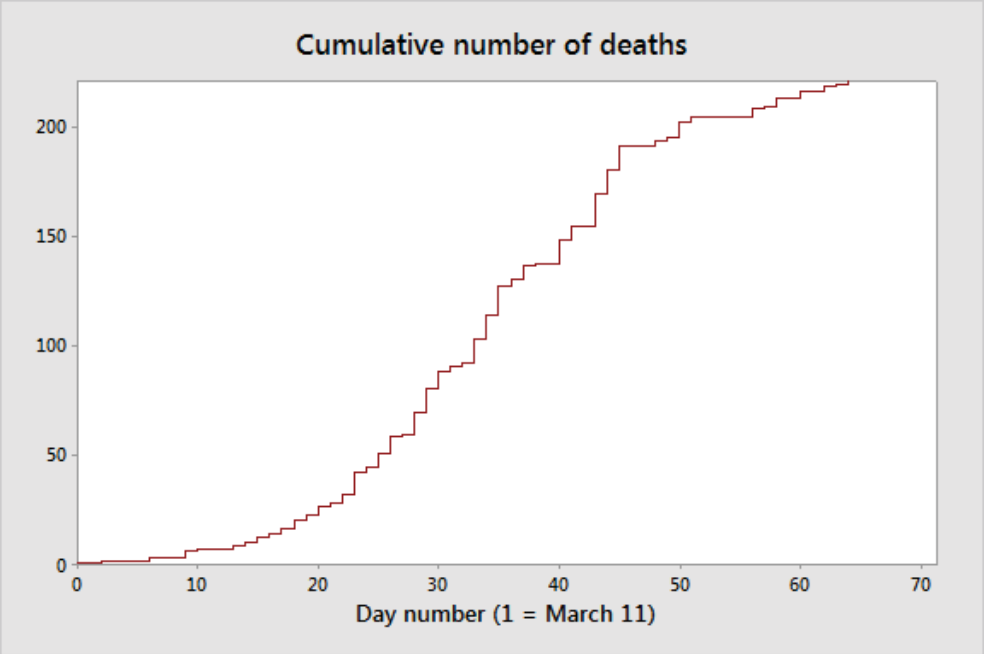

Figure 6: Cumulative number of deaths by COVID-19 in Norway from March 11 (day 1) to May 12 (day 63).

Let $N(t)$ be the number of deaths in the time interval $(0, t]$, defined in continuous time for $t > 0$ with time unit 'day'. Let D_i be the number of deaths reported on day number i. The D_i are hence interval counts for intervals $(0, 1], (1, 2], \ldots$, i.e.,

$$D_i = N(i) - N(i-1) \text{ for } i = 1, 2, \ldots.$$

Let $W(t) = E[N(t)]$ be the expected number of deaths in the interval $(0, t]$ and let the corresponding rate be defined as $w(t) = W'(t)$, for $t > 0$.

a) How would you estimate the function $W(t)$ nonparametrically for $0 < t \leq 63$ based on the data given in the table in Figure 5?

What can you say (roughly) about the function $w(t)$ based on the plot in Figure 6? How can the behaviour of $w(t)$ be interpreted in terms of deaths caused by COVID-19?

Let the $N(t)$ defined above be modeled by a nonhomogenous Poisson process (NHPP) with cumulative intensity function $W(t; \theta)$ and intensity function $w(t; \theta)$, where θ is an unknown parameter (which may be a vector). Suppose we have observed the D_i for $i = 1, 2, \ldots, r$ for some $r \geq 1$, with observations denoted $d_1, d_2, \ldots, d_r$.

b) Show how the properties of NHPPs lead to the following likelihood function for the observations $d_1, \ldots, d_r$:

$$L(\theta) = \left\{ \prod_{i=1}^{r} \frac{[W(i;\theta) - W(i-1;\theta)]^{d_i}}{d_i!} \right\} e^{-W(r;\theta)}. \tag{1}$$

Let now $\theta = (\alpha, \beta)$ and let the intensity function be given by

$$w(t;\alpha,\beta) = e^{\alpha+\beta t} \text{ for } t > 0$$

for real parameters α, β.

c) Show that the cumulative intensity function now can be written

$$W(t;\alpha,\beta) = \frac{e^{\alpha}}{\beta}\left(e^{\beta t} - 1\right) \text{ for } t > 0 \text{ and } \beta \neq 0.$$

What is the expression for $W(t;\alpha,\beta)$ if $\beta = 0$?

Use (1) to show that the log-likelihood of data $d_1, d_2, \ldots, d_r$ can be written

$$\ell(\alpha,\beta) = \left(\alpha + \ln\frac{1-e^{-\beta}}{\beta}\right)\sum_{i=1}^{r} d_i + \beta\sum_{i=1}^{r} i d_i - \sum_{i=1}^{r}\ln d_i! - \frac{e^{\alpha}}{\beta}(e^{\beta r} - 1). \tag{2}$$

Show that the maximum likelihood estimator of α when β is known, is given by

$$\hat{\alpha}(\beta) = \ln\left(\frac{\beta\sum_{i=1}^{r} d_i}{e^{r\beta} - 1}\right).$$

d) Suppose we are at Friday April 3, which is day number 24 in the numbering of Figure 5. Letting $r = 24$ in the log-likelihood (2) and using the data for the first 24 days in Figure 5, it can be shown that (you are not asked to do this) the maximum likelihood estimate of β is

$$\hat{\beta} = 0.1242.$$

Use this to find $\hat{\alpha}$ and to estimate the expected number of new deaths in the 7 days following Friday April 3. Compare the prediction to the actual number of deaths observed in these 7 days according to the table in Figure 5.