COLUMBIA UNIVERSITY
Statistical Inference and Modeling
Graded Homework 2: solutions

Prof. Marco Avella
TA: Ian Kinsella
Fall 2019

## Exercise 1

1. Given the little information available, assuming random censoring seems like a natural and convenient assumption. We do not see an obvious pattern in the censoring. Perhaps some auxiliary information such as weight, age, sex ,etc could explain part of the censoring...

2. The Kaplain-Meier estimates seem to suggest that treatment B, the more aggresive treatment, is more effective.

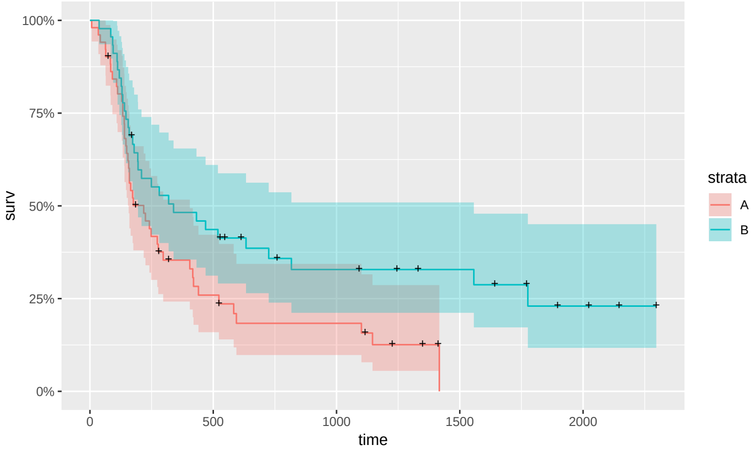

Figure 1: Kaplan-Meier estimates of the two treatments

3. The estimated coefficients do show a significant difference between the two groups. The sign of this difference matches what we saw in the preceeding part of this problem. Indeed, writing the survival function an exponential r.v. as $S(x) = e^{-\lambda x}$, the estimated cofficients for the two arms are $\hat{\lambda}_A = e^{-6.074}$ and $\hat{\lambda}_B = e^{-6.074-0.759}$.

```
Call:
survreg(formula = Surv(t, d) ~ arm, data = ncog, dist = "exp")
              Value Std. Error    z      p
(Intercept) 6.074      0.154 39.4 <2e-16
armB        0.759      0.237  3.2 0.0014

Scale fixed at 1

Exponential distribution
Loglik(model)= -539.9   Loglik(intercept only)= -545.1
```

```
Chisq= 10.41 on 1 degrees of freedom, p= 0.0013
Number of Newton-Raphson Iterations: 4
n= 96
```

4. The output given above also gives a p-value of 0.0013 using the likelihood ratio statistic which is approximately $\chi^2_1$. This provides evidence against $H_0$: "there is no difference between the treatments". The also agrees with the significant coefficient for arm B in the output. These conclusion depend on model assumptions (exponential parametric model) and asymptotic approximation to the distribution of the likelihood ratio statistic.

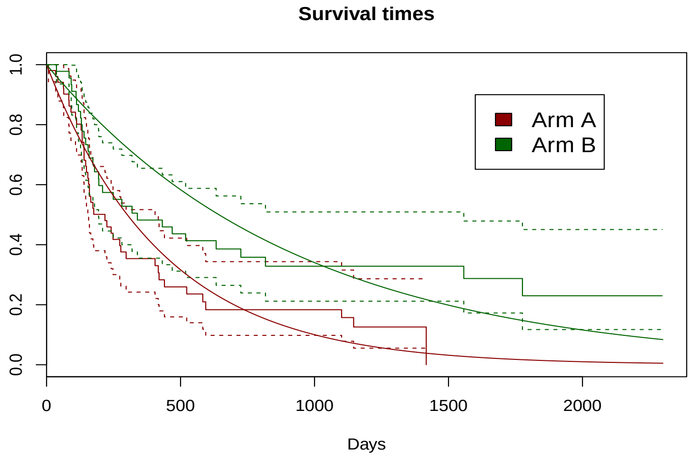

Figure 2: Checking the exponential fits

5. The visual check is rather good for the fitted survival function for treatment A. It is less satisfactory for treatment B because we see some small departures from the pointwise confidence intervals of the Kaplan-Meier counterpart around the end of the first year and for long survival times.

6. We see that the Weibull model does not seem to improve the fit of the exponential model. The added "scale" parameter is not significant and visually we cannot see any improvement in the estimated Weibull survival functions.

```
Call:
survreg(formula = Surv(t, d) ~ arm, data = ncog)
              Value Std. Error     z      p
(Intercept) 6.0387     0.1821 33.16 <2e-16
armB        0.7860     0.2789  2.82 0.0048
Log(scale)  0.1619     0.0921  1.76 0.0789

Scale= 1.18

Weibull distribution
```

```
Loglik(model)= -538.3   Loglik(intercept only)= -542.2
Chisq= 7.78 on 1 degrees of freedom, p= 0.0053
Number of Newton-Raphson Iterations: 5
n= 96
```

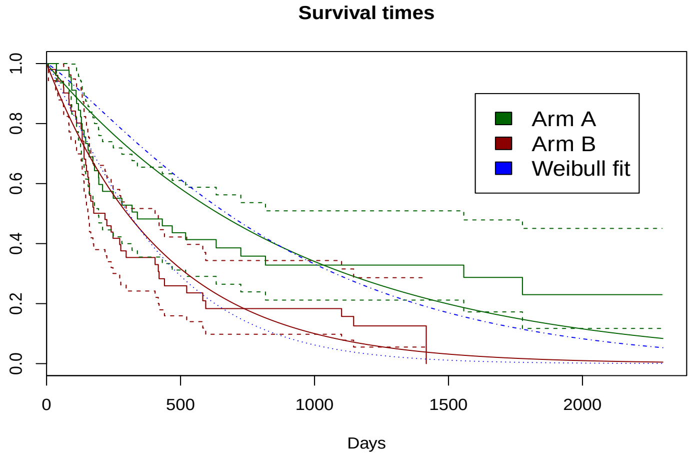

Figure 3: Checking the Weibull fits

## Exercise 2

1. We see large differences in the estimated covariance matrices due to the small sample size and the large number of missing entries for variables 1 and 5.

   (a) Complete case analysis:

   ```
   > cov(M,use="complete")
         x1       x2    x3       x4       x5
   x1 458.7 207.9000 135.4 180.8000 330.9000
   x2 207.9 337.3667 168.4 150.4667 372.5667
   x3 135.4 168.4000 160.4 110.0000 282.4000
   x4 180.8 150.4667 110.0 109.4667 211.8667
   x5 330.9 372.5667 282.4 211.8667 565.7667
   ```

   (b) Available case analysis:

   ```
   > cov(M,use="pairwise")
             x1        x2        x3        x4       x5
   x1 219.42424  95.15152  73.37879  93.66667 330.9000
   x2  95.15152 210.35968 127.02767 132.52569 236.5758
   x3  73.37879 127.02767 154.62451 131.91107 204.6364
   x4  93.66667 132.52569 131.91107 160.17391 173.1212
   x5 330.90000 236.57576 204.63636 173.12121 365.1515
   ```

   (c) Mean imputation:

```
> cov(mean.impute(M))
          x1        x2        x3        x4        x5
x1 109.71212  47.57576  36.68939  46.83333  75.02273
x2  47.57576 210.35968 127.02767 132.52569 118.28788
x3  36.68939 127.02767 154.62451 131.91107 102.31818
x4  46.83333 132.52569 131.91107 160.17391  86.56061
x5  75.02273 118.28788 102.31818  86.56061 182.57576
```

(d) Mean imputation with the bootstrap:

```
> bootimpS
         [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 98.84066  42.53268  33.41453  41.78593  64.77407
[2,] 42.53268 202.08684 122.41419 127.28666 108.11449
[3,] 33.41453 122.41419 149.99587 127.93255  93.98604
[4,] 41.78593 127.28666 127.93255 154.31263  79.98189
[5,] 64.77407 108.11449  93.98604  79.98189 167.02392
> impS
```

(e) The EM-algorithm:

```
> Mls(M)$sig
            x1        x2        x3       x4       x5
[1,] 322.13065  46.29316  94.09942 115.2921 196.2932
[2,]  46.29316 201.21361 121.50473 126.7637 167.3158
[3,]  94.09942 121.50473 147.90170 126.1758 194.0407
[4,] 115.29212 126.76371 126.17580 153.2098 136.8706
[5,] 196.29320 167.31583 194.04074 136.8706 336.3849
```

2. Taking the largest eigenvalue of the different estimated matrices gives different estimated eigenvalues. One can also onstruct a confidence interval for $\lambda_1$ using these different estimators of $\lambda_1$ in the

$$\left(\hat{\lambda}_1 - 1.96\frac{\sqrt{2}\hat{\lambda}_1}{\sqrt{n}}, \hat{\lambda}_1 + 1.96\frac{\sqrt{2}\hat{\lambda}_1}{\sqrt{n}}\right),$$

Applying naively this principle leads to the confidence intervals

(a) Complete case analysis: $(1133.0, 1459.7)$

(b) Available case analysis: $(807.5, 1040.4)$

(c) Mean imputation: $(483.1, 622.4)$

(d) Mean imputation with the bootstrap: $(256.5, 867.1)$

(e) The EM-algorithm: $(691.8, 891.2)$

Complete case analysis is useless for this data set because it only has 1 or 2 complete cases. Available case analysis is a slight improvement but the interval is very different from the one obtained with the EM algorithm. Mean imputaion gives the narrowest interval but is to be taken with caution since it underestimates the variance. The bootstrap confidence interval is the widest and therefore also the most conservative. With such small samples the validity of a normal appproximation is always questionable.

3. The confidence interval obtained using the full data set and the normal approximation is $(665.35, 708.63)$. This interval should be in principle more accurate as $\hat{\lambda}_1$ should have smaller bias and smaller variance. These two features will yield an interval that is both centered closer to the true population $\lambda_1$ and narrower.

   Interestingly the bootstrap method was the only one to provide a confidence interval that covers the interval $(665.35, 708.63)$ obtained with all the data. Furthermore if we use the same data and all the complete observations corresponding to the ones analyzed in the previous points, then we get the interval $(268.9, 286.4)$. Remarkably the bootstrap also covered this interval!

4. As derived in class, the log likelihood of the $i$th observation is given by

$$\begin{aligned}
\ell_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_i) &= -\tfrac{p}{2}\log 2\pi - \tfrac{1}{2}\log\det\boldsymbol{\Sigma} - \tfrac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \\
&= -\tfrac{p}{2}\log 2\pi - \tfrac{1}{2}\log\det\boldsymbol{\Sigma} - \tfrac{1}{2}\operatorname{tr}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}
\end{aligned}$$

   Thus

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\mu}}\ell_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_i) &= -\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \qquad (1) \\
\frac{\partial}{\partial\boldsymbol{\Sigma}}\ell_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_i) &= -\tfrac{1}{2}\boldsymbol{\Sigma}^{-1} + \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} \\
&= \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}((\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1} \qquad (2)
\end{aligned}$$

   Taking the conditional expectation of (1), summing over all $i$ and equalizing to zero shows that the estimating equation of $\mu$ is

$$\sum_{i=1}^{n}(\boldsymbol{\mu} - \hat{\mathbf{x}}_i) = 0.$$

   In order to derive the second estimating equation consider first the following calculations:

$$E_{\mathbf{x}_m|\mathbf{x}_o}[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T] = E_{\mathbf{x}_m|\mathbf{x}_o}\begin{bmatrix}(\mathbf{x}_{oi} - \boldsymbol{\mu}_{oi})(\mathbf{x}_{oi} - \boldsymbol{\mu}_{oi})^T & (\mathbf{x}_{oi} - \boldsymbol{\mu}_{oi})(\mathbf{x}_{mi} - \boldsymbol{\mu}_{mi})^T \\ (\mathbf{x}_{mi} - \boldsymbol{\mu}_{mi})(\mathbf{x}_{oi} - \boldsymbol{\mu}_{oi})^T & (\mathbf{x}_{mi} - \boldsymbol{\mu}_{mi})(\mathbf{x}_{mi} - \boldsymbol{\mu}_{mi})^T\end{bmatrix}$$

   with

$$\begin{aligned}
E_{\mathbf{x}_m|\mathbf{x}_o}[(\mathbf{x}_{oi} - \boldsymbol{\mu}_{oi})(\mathbf{x}_{oi} - \boldsymbol{\mu}_{oi})^T] &= (\hat{\mathbf{x}}_{oi} - \boldsymbol{\mu}_{oi})(\hat{\mathbf{x}}_{oi} - \boldsymbol{\mu}_{oi})^T \qquad (3) \\
E_{\mathbf{x}_m|\mathbf{x}_o}[(\mathbf{x}_{oi} - \boldsymbol{\mu}_{oi})(\mathbf{x}_{mi} - \boldsymbol{\mu}_{mi})^T] &= (\hat{\mathbf{x}}_{oi} - \boldsymbol{\mu}_{oi})(E_{\mathbf{x}_m|\mathbf{x}_o}[\mathbf{x}_{mi}] - \boldsymbol{\mu}_{mi})^T \\
&= (\hat{\mathbf{x}}_{oi} - \boldsymbol{\mu}_{oi})(\boldsymbol{\Sigma}_{imo}\boldsymbol{\Sigma}_{ioo}^{-1}(\hat{\mathbf{x}}_{oi} - \boldsymbol{\mu}_{oi}))^T \qquad (4)
\end{aligned}$$

and

$$\begin{aligned} E_{\mathbf{x}_m|\mathbf{x}_o}[(\mathbf{x}_{mi}-\boldsymbol{\mu}_{mi})(\mathbf{x}_{mi}-\boldsymbol{\mu}_{mi})^T] &= E_{\mathbf{x}_m|\mathbf{x}_o}[\mathbf{x}_{mi}\mathbf{x}_{mi}^T]-\boldsymbol{\mu}_{mi}\hat{\mathbf{x}}_{mi}^T-\hat{\mathbf{x}}_{mi}\boldsymbol{\mu}_{mi}^T-\boldsymbol{\mu}_{mi}\boldsymbol{\mu}_{mi}^T \\ &= (\hat{\mathbf{x}}_{mi}-\boldsymbol{\mu}_{mi})(\hat{\mathbf{x}}_{mi}-\boldsymbol{\mu}_{mi})^T \\ &\qquad +\boldsymbol{\Sigma}_{mmi}-\boldsymbol{\Sigma}_{moi}\boldsymbol{\Sigma}_{ooi}^{-1}\boldsymbol{\Sigma}_{omi} \qquad (5) \end{aligned}$$

where the last inequality comes from

$$\begin{aligned} E_{\mathbf{x}_m|\mathbf{x}_o}[\mathbf{x}_{mi}\mathbf{x}_{mi}^T] &= \boldsymbol{\Sigma}_{mmi}-\boldsymbol{\Sigma}_{moi}\boldsymbol{\Sigma}_{ooi}^{-1}\boldsymbol{\Sigma}_{omi}+E_{\mathbf{x}_m|\mathbf{x}_o}\mathbf{x}_{mi}E_{\mathbf{x}_m|\mathbf{x}_o}\mathbf{x}_{mi}^T \\ &= \boldsymbol{\Sigma}_{mmi}-\boldsymbol{\Sigma}_{moi}\boldsymbol{\Sigma}_{ooi}^{-1}\boldsymbol{\Sigma}_{omi}+\hat{\mathbf{x}}_{mi}\hat{\mathbf{x}}_{mi}^T \end{aligned}$$

Now taking the conditional expectation over (2) equalizing it to zero and using (3), (4) and (5) we have get

$$\sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-(\boldsymbol{\mu}-\hat{\mathbf{x}}_i)(\boldsymbol{\mu}-\hat{\mathbf{x}}_i)^T-\mathbf{C}_i)\boldsymbol{\Sigma}^{-1}=0.$$

Thus

$$\sum_{i=1}^{n}\Big(\boldsymbol{\Sigma}-(\boldsymbol{\mu}-\hat{\mathbf{x}}_i)(\boldsymbol{\mu}-\hat{\mathbf{x}}_i)^T-\mathbf{C}_i\Big)=0.$$

## Exercise 3

1. Each of the entries of the matrices above estimate the probability of going from state $i \in \{0,1\}$ at time $t$ to state $j \in \{0,1\}$ at time time $t+1$. (The two states should have been flipped i.e. $\{0,1\} = \{$“’no snow’, “snow”$\}$)

2. From the estimated transition matrix for New York we have

$$\mathbb{P}(X_{t+1}=1|X_t=1)=0.776$$

3. The equation defining the stationary distribution of $p_0$ is

$$0.861p_0+0.011(1-p_0)=p_0 \Longleftrightarrow p_0=\frac{11}{150}.$$

Therefore the estimated long-term probability of observing a snowday in Medford is approximately 0.92.

4. This amounts to testing $H_0 : p_{00} = p_{11}$. Using the asymptotic nomality of $(\hat{p}_{00}, \hat{p}_{11})$ and the asymptotic independence of their components, we have the following approximation under the null hypothesis

$$\hat{p}_{00}-\hat{p}_{11} \approx N(0, p_{00}(1-p_{00})/n_{0\cdot}+p_{11}(1-p_{11})/n_{1\cdot}).$$

This can be used a test statistics for $H_0$.

5. See R code.

6. The likelihood ratio statistic gives strong evidence against the first order model.

**Exercise 4**

1.
$$\begin{aligned} f(n_A, n_C, n_G, n_T; \theta) &= \mathbb{P}(N_A = n_A, N_C = n_C, N_G = n_G, N_T = n_T) \\ &= \frac{n!}{n_A! n_C! n_G! n_T!}(1-\theta)^{n_A}(\theta - \theta^2)^{n_C}(\theta^2 - \theta^3)^{n_G}(\theta^3)^{n_T} \end{aligned}$$

2. Since
$$\frac{\partial \log f(n_A, n_C, n_G, n_T; \theta)}{\partial \theta} = -\frac{n_A}{1-\theta} + \frac{(1-2\theta)n_C}{\theta - \theta^2} + \frac{(2\theta - 3\theta^2)n_G}{\theta^2 - \theta^3} + \frac{3n_T}{\theta}$$
the MLE is defined by setting to zero the above equation. Multiplying by $\theta(1-\theta)$ leads to
$$(n_A + 2n_C + 3n_G + 3n_T)\hat{\theta} = n_C + 2n_G + 3n_T$$
Show that the MLE of $\theta$ is
$$\hat{\theta} = \frac{N_C + 2N_G + 3N_T}{N_A + 2N_C + 3N_G + 3N_T}$$

3.
$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, \theta(1-\theta)/(1+\theta+\theta^2)$$

4. We want to have
$$\mathbb{E}[T] = n(a_A(1-\theta) + a_C(\theta - \theta^2) + a_G(\theta^2 - \theta^3) + a_T\theta^3) = \theta.$$
We can therefore pick $a_A = 0$ and $a_C = a_G = a_T = 1/n$.

5. Note that with the above choices of constants $a_A, a_C, a_G, a_T$ we have that $T = (n - N_A)/n$ and hence
$$\operatorname{var}(T) = \frac{1}{n^2}\operatorname{var}(n - N_A) = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}.$$
The asymptotic relative efficiency is
$$\frac{1}{n}\frac{\theta(1-\theta)}{1+\theta+\theta^2} \Big/ \Big(\frac{1}{n}\theta(1-\theta)\Big) = 1/\Big(1+\theta+\theta^2\Big).$$

6. Without any dependence on $\theta$ the MLE is $(\hat{p}_A, \hat{p}_C, \hat{p}_G, \hat{p}_T) = (\frac{N_A}{n}, \frac{N_C}{n}, \frac{N_G}{n}, \frac{N_T}{n})$ while the MLE using point 2 is $(1-\hat{\theta}, \hat{\theta}(1-\hat{\theta}), \hat{\theta}^2(1-\hat{\theta}), \hat{\theta}^3)$. Assuming $a_A = 0$ and $a_C = a_G = a_T = 1/n$, $T = (N_C + N_G + N_T)/n$ which is an estimator of $1 - p_A$ corresponding exactly to $1 - \hat{p_A}$.

7. Let $\mathbf{p}(\theta) = (1-\theta, \theta-\theta^2, \theta^2-\theta^3, \theta^3)^T$. Using the asymptotic normality of $\hat{\mathbf{p}} = (\hat{p}_A, \hat{p}_C, \hat{p}_G, \hat{p}_T)^T$ we can contruct the Wald statistics for a fixed value of $\theta$ as
$$W_n = n(\hat{\mathbf{p}} - \mathbf{p}(\theta))^T \hat{\mathbf{V}}^{-1}(\hat{\mathbf{p}} - \mathbf{p}(\theta)),$$
where
$$\hat{\mathbf{V}}_{ij} = \begin{cases} \hat{p}_i(1-\hat{p}_i)/n, & i = j \in \{A, C, G, T\}, \\ -\hat{p}_i\hat{p}_j/n, & i \neq j \in \{A, C, G, T\}. \end{cases}$$
Under $H_0 : \mathbf{p} = \mathbf{p}(\theta)$ we have that $W_n \overset{D}{\to} \chi^2_4$

**Exercise 5** (Optional bonus question)

1. There is a significant negative slope paramater suggesting a negative trend of this time series. The residuals are clearly correlated as the signs of the residual at time $t$ strongly influences the sign of the residuals in the next time periods.

2. The ACF and PACF confirm the presence of a time dependence structure. In particular the ACF tails off while the PACF cuts off.

3. The AIC suggests that an AR(2) fitted to the residuals is a better model.

4. The AIC suggest AR(1) and ARMA(1,2) fits for the the residuals of the linear model. The BIC suggest ARMA(1,1) and ARMA(1,2).