Exam 2

Xinyu Zhang

Instructions

- Please hand in a hard copy by **Tuesday, December 11**.
- This is a open resource exam, but you are not allowed to ask post exam questions online.
- You are not allowed to collaborate with classmates and/or people outside of class.
- Please circle or highlight your final answer.
- The total possible point is 50.

Violation of this agreement will result in an **F** on this exam and it will be averaged in as a 0%.

---

1. Under the exponential model, we assume T follows an exponential distribution with parameter λ and derived the maximum likelihood for λ to be

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \Delta_i}{\sum_{i=1} t_i},$$

where Δ_i's is the censoring indicator and t_i's are the observed survival times. Now let's suppose each copy of $\{\Delta_i, t_i\}$ is associated with a weight w_i (this weight could represent sampling weights, propensity score, or counts, etc.).

   a. (5 points) Modify the likelihood presented in note 3 and derived the weighted maximum likelihood estimator for λ. We will denote this $\hat{\lambda}_w$.

If we just multiply each term of the likelihood function with its weight, then the weight will become 0 when we take the derivative of the log likelihood. Therefore, we raise the power of each term of the likelihood function to its weight and continue from there.

$$L(\lambda) = \prod_{i=1}^{n} \left[\lambda e^{-\lambda t_i}\right]^{\Delta_i \cdot w_i} \cdot \left[e^{-\lambda t_i}\right]^{(1-\Delta_i)\cdot w_i} = \prod_{i=1}^{n} \lambda^{\Delta_i \cdot w_i} \cdot e^{-\lambda t_i \cdot w_i}$$

$$\log L(\lambda) = \ell(\lambda) = \log(\lambda) \left(\sum_{i=1}^{n} \Delta_i \cdot w_i\right) - \lambda \sum_{i=1}^{n} t_i \cdot w_i$$

$$\frac{\mathrm{d} \log L(\lambda)}{\mathrm{d}\lambda} = \ell'(\lambda) = 0 \text{ gives } \hat{\lambda}_w = \frac{\sum_{i=1}^{n} \Delta_i \cdot w_i}{\sum_{i=1}^{n} t_i \cdot w_i}$$

   b. (5 points) Use `los` (length of stay) as the weight in *WHAS100*. Apply `survreg` to *WHAS100* to compute $\hat{\lambda}_w$.

```
> fm <- Surv(lenfol, fstat) ~ 1
> fit.aft.weighted <- survreg(fm, data = whas100, dist="exp", weight=los)
>
> summary(fit.aft.weighted)
```

```
Call:
survreg(formula = fm, data = whas100, weights = los, dist = "exp")
            Value Std. Error   z p
(Intercept)     8     0.0521 153 0

Scale fixed at 1

Exponential distribution
Loglik(model)= -3312.6   Loglik(intercept only)= -3312.6
Number of Newton-Raphson Iterations: 4
n= 100
```

We see that the intercept is 8 which is $-\log(\lambda_w)$, so $\lambda_w = e^{-8} = 3.3546 \times 10^{-4}$.

c. (5 points) Use the derivation in 1a and *WHAS100* to compute $\hat{\lambda}_w$.

```
> sum(whas100$fstat*whas100$los)/sum(whas100$lenfol*whas100$los)
```

```
[1] 0.0003349529
```

3.3495×10^{-4} is very close to 3.3546×10^{-4}.

d. (5 points) Derive the *information* and the asymptotic variance for $\hat{\lambda}_w$.

$$\ell''(\lambda_w) = -\frac{1}{\lambda_w^2} \sum_{i=1}^{n} \Delta_i \cdot w_i$$

$$Var(\hat{\lambda}_w) \approx \frac{\hat{\lambda}_w^2}{\sum_{i=1}^{n} \Delta_i \cdot w_i}$$

2. In exam 1, we have investigated different methods to compare two survival curves. Another way to compare two survival curves is to fit a Cox model using the group indicator as the covariate. Use the complete *WHAS100* dataset and gender as the group indicator for the following questions.

a. (5 points) Fit a Cox model and print the summary using gender as the covariate.

```
> fm <- Surv(lenfol, fstat) ~ gender
> fit.cox.gender <- coxph(fm, data = whas100)
>
> summary(fit.cox.gender)
```

```
Call:
coxph(formula = fm, data = whas100)

  n= 100, number of events= 51

         coef exp(coef) se(coef)     z Pr(>|z|)
gender 0.5548    1.7416   0.2824 1.965   0.0494 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
```

```
gender     1.742     0.5742     1.001     3.029

Concordance= 0.565  (se = 0.035 )
Rsquare= 0.037   (max possible= 0.985 )
Likelihood ratio test= 3.75  on 1 df,   p=0.05293
Wald test            = 3.86  on 1 df,   p=0.04945
Score (logrank) test = 3.96  on 1 df,   p=0.04665
```

b. (5 points) Interpret the estimated regression parameter ($\hat{\beta}$) in terms of hazard ratio.

The hazard ratio that compares male(gender=1) with female(gender=0) is 1.7416. For the change from female to male, the risk of death is expected to increase by 1.7416 times. The log of survival time is expected to decrease by 0.5548 (days).

c. (5 points) Use part 2a to test for the null hypothesis that there is no significant difference between the two survival curves.

We see that the p-value for b_{gender} is 0.0494 which is slightly smaller than α=0.05. Therefore we reject the H_0: b_{gender}=0 and conclude that there is a significant difference between male and female survival curves, but it is very marginal.

3. a. (5 points) Fit a Cox model with `age` as the (only) covariate. Plot the estimated cumulative hazard function for patients with `age = 50`, `age = 60` and `age = 70`.

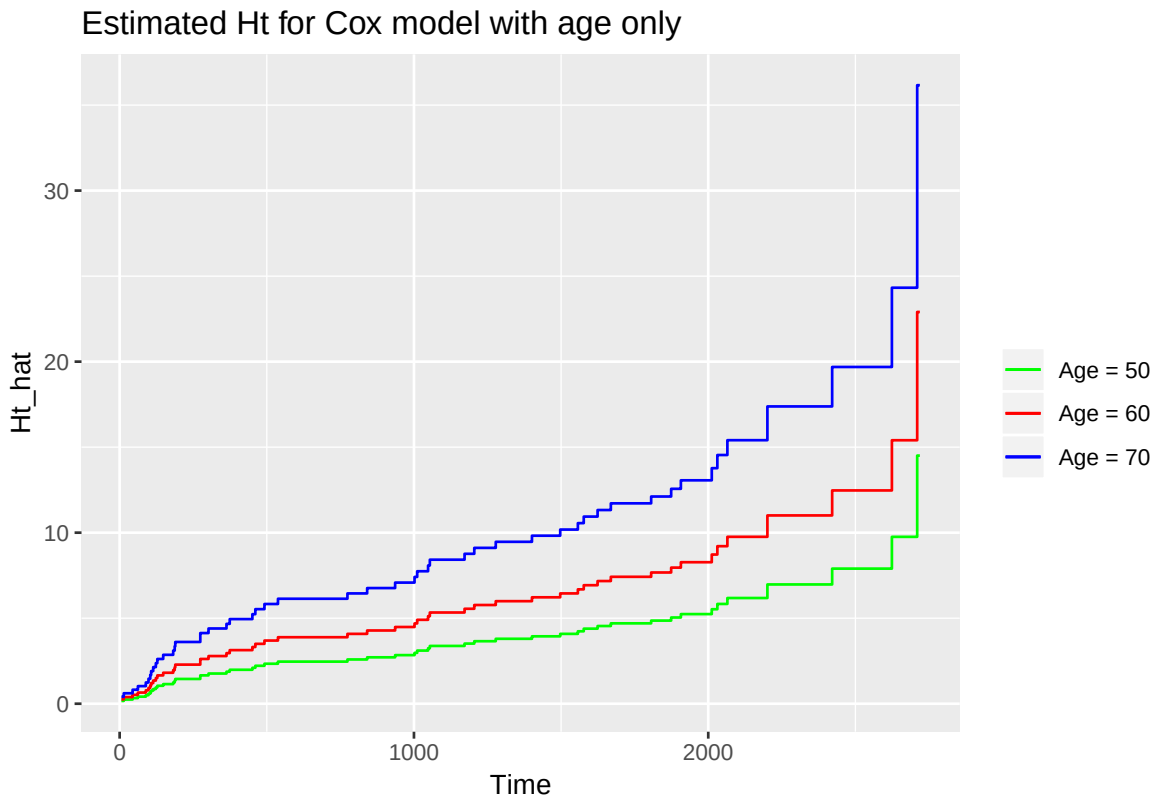
```
> fm <- Surv(lenfol, fstat) ~ age
> fit.cox.age <- coxph(fm, data = whas100)
>
> summary(fit.cox.age)
```

```
Call:
coxph(formula = fm, data = whas100)

  n= 100, number of events= 51

       coef exp(coef) se(coef)     z Pr(>|z|)
age 0.04567   1.04673  0.01195 3.822 0.000132 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    exp(coef) exp(-coef) lower .95 upper .95
age     1.047     0.9554     1.022     1.072

Concordance= 0.664  (se = 0.043 )
Rsquare= 0.159   (max possible= 0.985 )
Likelihood ratio test= 17.36  on 1 df,   p=3.09e-05
Wald test            = 14.61  on 1 df,   p=0.0001324
Score (logrank) test = 15.64  on 1 df,   p=7.675e-05
```

The coefficient of age is 0.04567.

```
> H.fit.cox.age=basehaz(fit.cox.age) %>%
+ mutate(H.50age=hazard*exp(0.04567*50),
+        H.60age=hazard*exp(0.04567*60),
+        H.70age=hazard*exp(0.04567*70)) %>%
```

```
+ select(time,H.50age,H.60age,H.70age)
>
> ggplot(data = H.fit.cox.age, aes(x = time)) +
+   geom_step(aes(y = H.50age, colour = "Age = 50")) +
+   geom_step(aes(y = H.60age, colour = "Age = 60")) +
+   geom_step(aes(y = H.70age, colour = "Age = 70")) +
+   scale_colour_manual("",breaks = c("Age = 50", "Age = 60", "Age = 70"),
+                       values = c("Age = 50"="green", "Age = 60"="red","Age = 70"="blue")) +
+   labs(title ="Estimated Ht for Cox model with age only", x = "Time", y = "Ht_hat")
```



b. (5 points) Fit another Cox model with `age` and `age^2` as covariates. Plot the estimated cumulative hazard function for patients with `age = 50`, `age = 60` and `age = 70`.

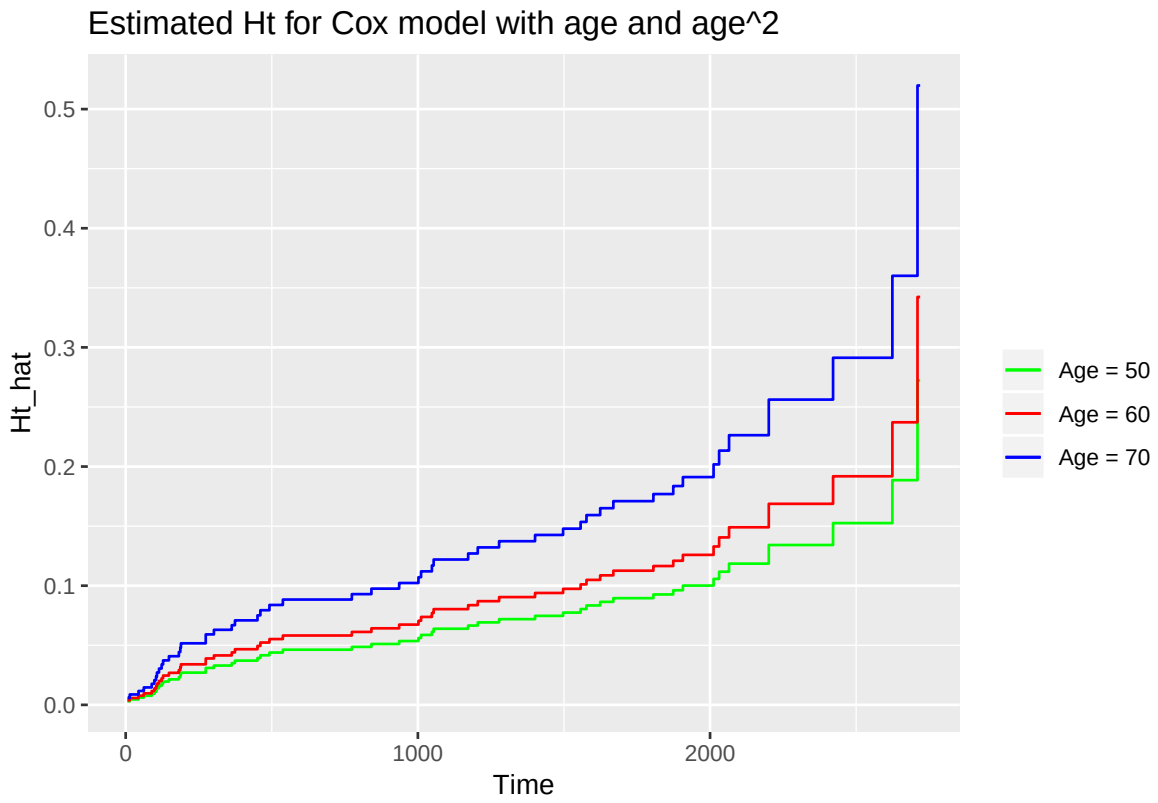
```
> fm2 <- Surv(lenfol, fstat) ~ age + I(age^2)
> fit.cox.age2 <- coxph(fm2, data = whas100)
>
> summary(fit.cox.age2)
```

```
Call:
coxph(formula = fm2, data = whas100)

  n= 100, number of events= 51

               coef  exp(coef)   se(coef)      z Pr(>|z|)
age      -0.0808322  0.9223484  0.0922260 -0.876    0.381
I(age^2)  0.0009431  1.0009435  0.0006884  1.370    0.171
```

```
          exp(coef) exp(-coef) lower .95 upper .95
age          0.9223     1.0842    0.7698     1.105
I(age^2)     1.0009     0.9991    0.9996     1.002

Concordance= 0.666  (se = 0.043 )
Rsquare= 0.174   (max possible= 0.985 )
Likelihood ratio test= 19.09  on 2 df,   p=7.161e-05
Wald test            = 19.13  on 2 df,   p=7.016e-05
Score (logrank) test = 21.23  on 2 df,   p=2.448e-05
```

The coefficient of age is -0.0808322 and age^2 is 0.0009431.

```
> H.fit.cox.age2=basehaz(fit.cox.age2) %>%
+ mutate(H.50age=hazard*exp(-0.0808322*50+0.0009431*50^2),
+        H.60age=hazard*exp(-0.0808322*60+0.0009431*60^2),
+        H.70age=hazard*exp(-0.0808322*70+0.0009431*70^2)) %>%
+ select(time,H.50age,H.60age,H.70age)
>
> ggplot(data = H.fit.cox.age2, aes(x = time)) +
+   geom_step(aes(y = H.50age, colour = "Age = 50")) +
+   geom_step(aes(y = H.60age, colour = "Age = 60")) +
+   geom_step(aes(y = H.70age, colour = "Age = 70")) +
+   scale_colour_manual("",breaks = c("Age = 50", "Age = 60", "Age = 70"),
+                       values = c("Age = 50"="green", "Age = 60"="red","Age = 70"="blue")) +
+   labs(title ="Estimated Ht for Cox model with age and age^2", x = "Time", y = "Ht_hat")
```



c. (5 points) Interpret the results in 3a and 3b.

For both models, the Nelson-Aalen estimator baseline hazard functions increase rapidly at the very beginning and the very end, but slowdown in the middle portion. The range of the hazard function for the first model is almost 70 times larger than the second one, so the two models provide significantly different hazard function.

```
> 1 - pchisq(2 * sum(fit.cox.age2$loglik - fit.cox.age$loglik), 1)
```

```
[1] 0.1888298
```

The partial likelihood ratio test gives a p-value of 0.1888 larger than α=0.05, so we cannot reject the H_0: b_{age^2}=0 and conclude that age^2 is not significant enough to include in the model. Thus, we should select the first model and use its hazard function.