STAT 639V - Assignment 3

Carson Stacy

11/03/2021

1. Fit a Cox proportional hazard model for the lung cancer data in survival package: $Time \sim Sex + Age$

```r
library(survival)
# add survival object column
lung$SurvObj <- with(lung, Surv(time, status == 2))

# # change male to 0, female to 1.
# lung$sex01 <- ifelse(lung$sex == 1, 0, 1)

# fit the Cox proportional hazard model for the data
coxreg <- coxph(SurvObj ~ sex + age, lung)
```

(a) Interpret the estimated coefficients.

```r
summary(coxreg)
```

```
## Call:
## coxph(formula = SurvObj ~ sex + age, data = lung)
##
##   n= 228, number of events= 165
##
##          coef exp(coef)  se(coef)      z Pr(>|z|)
## sex -0.513219  0.598566  0.167458 -3.065  0.00218 **
## age  0.017045  1.017191  0.009223  1.848  0.06459 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## sex    0.5986     1.6707    0.4311    0.8311
## age    1.0172     0.9831    0.9990    1.0357
##
## Concordance= 0.603  (se = 0.025 )
## Likelihood ratio test= 14.12  on 2 df,   p=9e-04
## Wald test            = 13.47  on 2 df,   p=0.001
## Score (logrank) test = 13.72  on 2 df,   p=0.001
```

The Cox model fit finds the effect of sex on survival to be significant with a p-value of 0.0022, while age is not significant at an $\alpha = 0.05$ in the model with a p-value of 0.065. The beta coefficient of -0.513219 for sex indicates that, all else equal, females in the study had lower risk of death in the study. This is because the category assigned a higher value (in this case, Female where Female=2 and Male=1) has lower risk than the baseline group due to the negative value of the beta coefficient. The magnitude of the coefficient depends on how much of a difference of there is between male and female survival. The exponentiated coefficient exp(coef) for sex is 0.599 which is the hazard ratio, which means being female corresponds to a hazard reduction of about 40%.

On the other hand, the positive coefficient of the numerical variable age corresponds to an increase in risk as age increases. The magnitude of the coefficient relates to the amount risk increases per unit increase in age. The exponentiated coefficient of sex is 1.02 which indicates a 2% increase in hazard per year of age above baseline. However, this parameter is not found to be significant in the model estimate, so caution is needed in considering this covariate.

(b) Compute the point estimate and 95% interval estimate for the hazard ratio between female and male. (20 points)

The point estimate for the hazard ratio for sex in this model is included in the summary above, with $HR = 0.5986$, 95% $CI = [0.4311, 0.8311]$. This means approximately a 40% lower hazard for female compared to male patients, with the true hazard corresponding to between a 27% to 57% reduction in hazard at a confidence level of 95%.

We confirm these values below:

$$Male = 1; Female = 2$$

$$HR = exp((Female - Male)\hat{\beta}) = exp((2-1)\hat{\beta}) = exp(\hat{\beta})$$

If $\hat{\beta} = -0.513219$ for sex, then we have:

$$HR = exp(-0.513219) \approx \boxed{0.599} = HR$$

To solve for the 95% confidence interval for the hazard ratio of sex:

$$exp((Female - Male)\hat{\beta} \pm z_{\alpha/2}(Female - Male)\hat{SE}(\hat{\beta}))$$

$$CI_{95\%} = exp(\hat{\beta} \pm z_{\alpha/2}\hat{SE}(\hat{\beta}))$$

With $\hat{\beta} = -0.513219$ and $\hat{SE}(\hat{\beta}) = 0.167458$ from the summary above, we need to find the z-score for the 95% interval, which has an alpha of 0.05:

```
qnorm(0.025)
```

```
## [1] -1.959964
```

Solving the equation above, we have:

$$CI_{95\%} = exp(-0.513219 \pm -1.959964 \times 0.167458)$$

Lower bound of confidence interval:

$$CI_{95\%} = exp(-0.513219 + (-1.959964 \times 0.167458)) = 0.4310933$$

Upper bound of confidence interval:

$$CI_{95\%} = exp(-0.513219 - (-1.959964 \times 0.167458)) = 0.8310982$$

These calculations show that the summary HR estimate and confidence intervals are correct.