ST525 HW 8

Nora Quick

Question 1

Part (a)

$h(t|status) = h_0(t|status)exp(\beta_1 covariate)$

Part (b)

$PL(\beta) = exp(status^T\beta)/\Sigma exp(status^T\beta)$

Question 2

Part (a1)

No, I do not think that all of the types can be included in the Cox PH model. I think this because it is a comparison between the effects of similar things and in this case it appears that while there are groups within the types (1&3 and 2&4) I don't believe the two sets can be compared together.

Part (b1)

$h(time|status) = h_0(time|status)exp(\beta_1 Type2 + \beta_2 Type3 + \beta_3 Type4 + \beta_4 Trt + \beta_5 Karno + \beta_6 Age + \beta_7 Prior)$

Part (c1)

A major key assumption is that all the types can be compared together (which we discussed on part a). We also need to assume that all were diagnosed at the same time and that everything is proportional.

Part (a2)

Yes, it appears that there were a few covariates effect the survival times. Trt, Diagtime, Prior therapy, and Type2 all have statistically significant p-values and relatively large (positive) estimates.

Part (b2)

No, patients in standard chemotherapy do not have a longer expected survival time compared to the test chemotherapy. I say this because of the conclusions drawn in lab. The estimate for standard chemotherapy shows that the risk of death is 1.06 times higher for patients with standard treatement.

Part (c2)

Yes, it appears that patients with small cell type have a longer expected survival rate compared to Type1. The estimate appears to indicate that it has a 0.66 the risk factor of Type1 indicating it has less of a risk factor and, therefore, we can conclude that it has a longer expected survival rate compared to squamous type.

Part (d2)

Type2 appears to have longer survival rate than Type1 and Type4 due to the estimate of 0.66. Additionally it has a statistically significant p-value indicating that it has an effect on the survival time/risk factor.

Part (e2)

Diagtime is also a statistically significant values with a p-value of 0.0008 indicating that is has an effect on survival time/risk factor. It also appears to indicate that Diagtime has a longer survival rate with an estimate of 0.72.

Question 3

```r
aids <- read.table('~/AIDS.txt', header = FALSE)
head(aids)
```

```
##   V1  V2 V3  V4 V5 V6 V7 V8 V9 V10 V11 V12 V13   V14 V15 V16
## 1  1 189  0 189  0  0  1  1  1   1   1   0 100 169.0  39  34
## 2  2 287  0 287  0  0  1  1  2   2   1   0  90 149.5  15  34
## 3  3 242  0 242  0  1  2  0  1   1   1   1 100  23.5   9  20
## 4  4 199  0 199  0  0  1  1  1   1   1   0  90  46.0  53  48
## 5  5 286  0 286  0  1  2  0  1   1   3   0  90  10.0  12  46
## 6  6 285  0 285  0  1  2  0  1   1   1   0  70   0.0  24  51
```

```r
aidsIDV <- aids[aids$V6 != 0, ]
head(aidsIDV)
```

```
##    V1  V2 V3  V4 V5 V6 V7 V8 V9 V10 V11 V12 V13   V14 V15 V16
## 3   3 242  0 242  0  1  2  0  1   1   1   1 100  23.5   9  20
## 5   5 286  0 286  0  1  2  0  1   1   3   0  90  10.0  12  46
## 6   6 285  0 285  0  1  2  0  1   1   1   0  70   0.0  24  51
## 8   8 285  0 285  0  1  2  1  1   2   3   0  80 117.5  24  40
## 11 11 334  0 334  0  1  2  0  2   3   1   0  90   7.5  62  38
## 12 12 285  0 285  0  1  2  1  2   2   3   0  90  43.5  19  40
```

```r
aidsCont <- aids[aids$V6 == 0, ]
head(aidsCont)
```

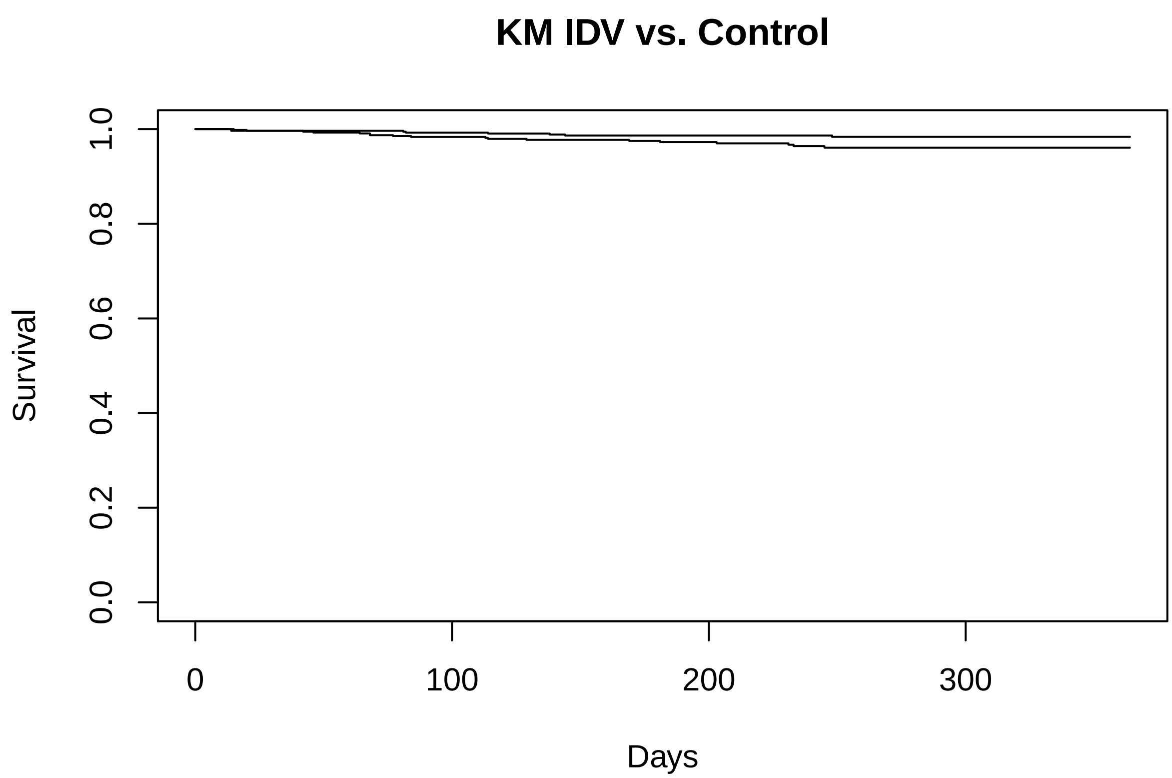
```
##    V1  V2 V3  V4 V5 V6 V7 V8 V9 V10 V11 V12 V13   V14 V15 V16
## 1   1 189  0 189  0  0  1  1  1   1   1   0 100 169.0  39  34
## 2   2 287  0 287  0  0  1  1  2   2   1   0  90 149.5  15  34
```

```
## 4    4 199   0 199   0   0   1   1   1    1    1    0   90   46.0   53   48
## 7    7 270   0 270   0   0   1   1   1    2    1    0  100   54.5    6   51
## 9    9 276   0 276   0   0   1   1   1    1    1    0  100   95.0    7   34
## 10  10 306   0 306   0   0   1   1   1    1    1    0   90   71.0    7   38
```

Part (a)

```r
KM <- survfit(Surv(V2, V5) ~ V6, data = aids, conf.type="none")

plot(KM, main = 'KM IDV vs. Control', xlab = 'Days',  ylab = 'Survival')
```




```r
#KM1 <- survfit(Surv(V2, V5) ~ V6, data = aidsIDV, conf.type="none")

#KM0 <- survfit(Surv(V2, V5) ~ V6, data = aidsCont, conf.type="none")

#plot(KM1, main = 'KM IDV', xlab = 'Days',  ylab = 'Survival')

#plot(KM0, main = 'KM Control', xlab = 'Days',  ylab = 'Survival')
```

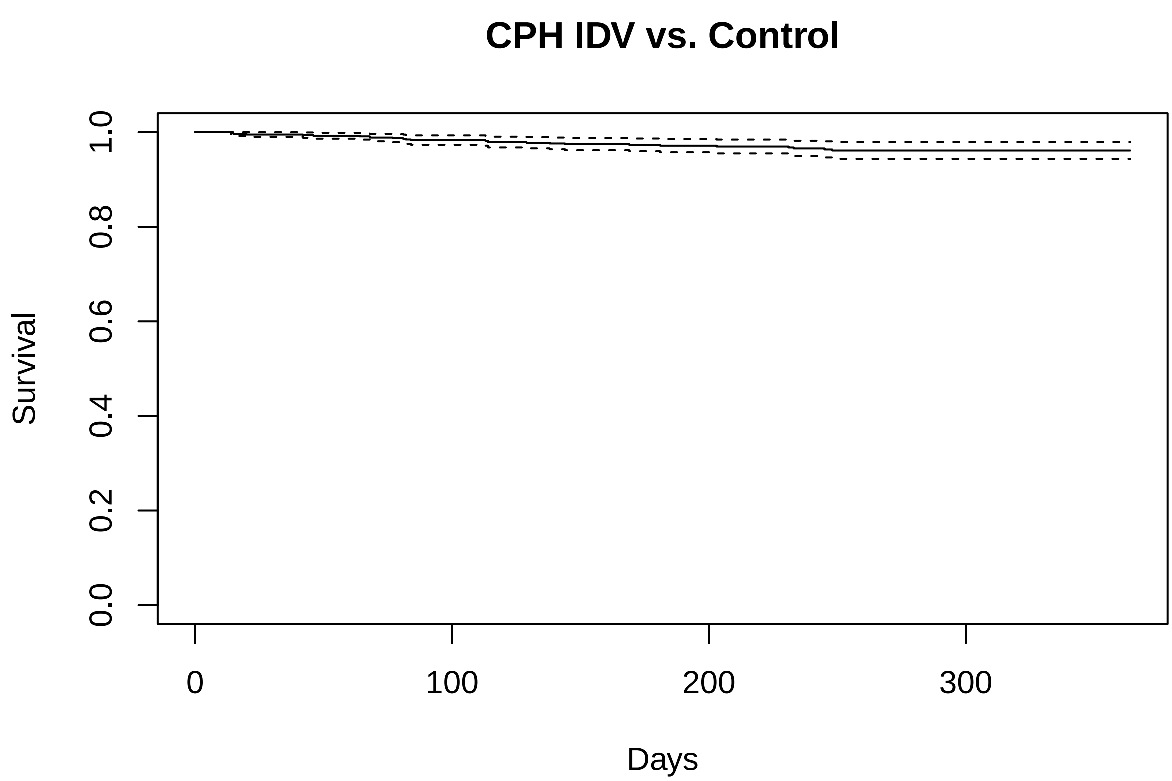
Part (b)

```r
CPH1 <- coxph(formula = Surv(V2, V5) ~ V6, data = aids)

summary(CPH1)
```

```
## Call:
## coxph(formula = Surv(V2, V5) ~ V6, data = aids)
##
##   n= 1151, number of events= 26
##
##        coef exp(coef) se(coef)      z Pr(>|z|)
## V6 -0.8520    0.4265   0.4250 -2.005    0.045 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##    exp(coef) exp(-coef) lower .95 upper .95
## V6    0.4265      2.344    0.1854    0.9811
##
## Concordance= 0.598  (se = 0.046 )
## Likelihood ratio test= 4.37  on 1 df,   p=0.04
## Wald test            = 4.02  on 1 df,   p=0.04
## Score (logrank) test = 4.27  on 1 df,   p=0.04
```

```r
plot(survfit(CPH1), main = 'CPH IDV vs. Control', xlab = 'Days',  ylab = 'Survival')
```



Part (c)

Based on the graphs I got it appears that the two plots differ only slightly and I believe this is mostly because of the singlular covariant. It does appear that the Cox PH model has a bigger difference between the two treatements and a clear mean.

As for what I can say about the proportional hazard assumption I'd say it appears that it is relatively proportional.

Part (d)

```r
CPH10 <- coxph(formula = Surv(V2, V5) ~ V10, data = aids)

summary(CPH10)
```

```
## Call:
## coxph(formula = Surv(V2, V5) ~ V10, data = aids)
##
##   n= 1151, number of events= 26
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## V10 0.2366    1.2669   0.2046 1.156    0.248
##
##     exp(coef) exp(-coef) lower .95 upper .95
## V10     1.267     0.7893    0.8484     1.892
##
## Concordance= 0.555  (se = 0.056 )
## Likelihood ratio test= 1.25  on 1 df,   p=0.3
## Wald test            = 1.34  on 1 df,   p=0.2
## Score (logrank) test = 1.35  on 1 df,   p=0.2
```

There is no difference between the races when using coxph? This is not shown in lab or elsewhere and I cannot produce something other than this.

The regression coefficient of race shows that it is not a statistically significant factor and that there is a 1.26 times higher risk factor for races other than white.

```r
KM2 <- survfit(Surv(V2, V5) ~ V10, data = aids, conf.type="none")
summary(KM2)
```

```
## Call: survfit(formula = Surv(V2, V5) ~ V10, data = aids, conf.type = "none")
##
##                  V10=1
##  time n.risk n.event survival std.err
##    15    586       1    0.998 0.00171
##    20    583       1    0.997 0.00241
##    46    575       1    0.995 0.00297
##    68    563       1    0.993 0.00345
##    82    555       1    0.991 0.00388
##    84    553       1    0.989 0.00427
##   114    529       1    0.988 0.00465
##   129    516       1    0.986 0.00502
```

```
##   138    502       1    0.984 0.00538
##   144    496       1    0.982 0.00572
##   233    375       1    0.979 0.00628
##
##                  V10=2
##  time n.risk n.event survival std.err
##    14    325       1    0.997 0.00307
##    42    317       1    0.994 0.00439
##    77    298       1    0.990 0.00549
##   114    273       1    0.987 0.00656
##   169    234       1    0.983 0.00777
##   181    222       1    0.978 0.00891
##   231    177       1    0.973 0.01043
##   245    156       1    0.966 0.01209
##   248    155       1    0.960 0.01352
##
##                  V10=3
##  time n.risk n.event survival std.err
##    68    181       1    0.994 0.00551
##    81    176       1    0.989 0.00786
##   113    167       1    0.983 0.00979
##   203    135       1    0.976 0.01213
##
##                  V10=4
##  time n.risk n.event survival std.err
##    14     14       1    0.929  0.0688
##    64     12       1    0.851  0.0973
##
##                  V10=5
##      time n.risk n.event survival std.err
```

Based on THIS summary, however, it appears that the surival rates between “blacks” and “Hispanics” are relatively the same.

Part (e)

```
CPH <- coxph(formula = Surv(V2, V5) ~ V6 + V8 + V9 + V11 + V12 + V13 + V15 + V16, data = aids)

summary(CPH)
```

```
## Call:
## coxph(formula = Surv(V2, V5) ~ V6 + V8 + V9 + V11 + V12 + V13 +
##     V15 + V16, data = aids)
##
##   n= 1151, number of events= 26
##
##          coef exp(coef)  se(coef)      z Pr(>|z|)
## V6  -0.825866  0.437856  0.427781 -1.931 0.053535 .
## V8  -1.151292  0.316228  0.447772 -2.571 0.010136 *
## V9   0.614486  1.848706  0.511674  1.201 0.229777
## V11 -0.046773  0.954304  0.264537 -0.177 0.859658
```

```
## V12  1.033174  2.809969  1.049601  0.984 0.324944
## V13 -0.095055  0.909323  0.023376 -4.066 4.78e-05 ***
## V15 -0.013971  0.986127  0.009729 -1.436 0.151000
## V16  0.073156  1.075898  0.019527  3.746 0.000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## V6     0.4379     2.2839    0.1893    1.0126
## V8     0.3162     3.1623    0.1315    0.7606
## V9     1.8487     0.5409    0.6782    5.0397
## V11    0.9543     1.0479    0.5682    1.6027
## V12    2.8100     0.3559    0.3592   21.9848
## V13    0.9093     1.0997    0.8686    0.9520
## V15    0.9861     1.0141    0.9675    1.0051
## V16    1.0759     0.9295    1.0355    1.1179
##
## Concordance= 0.88  (se = 0.023 )
## Likelihood ratio test= 50.97  on 8 df,   p=3e-08
## Wald test            = 51.44  on 8 df,   p=2e-08
## Score (logrank) test = 57.4  on 8 df,   p=2e-09
```

Based on the output above there appear to be two covarate variables that are statistically significant. The Karnofsky performance and the age at enrollment. Of course, those two make sense since one indicates how sick someone is (the sicker they are the sooner they are likely to die) and the older the enrollment (the older they are at enrollment the sooner they are likely to die) are reasonable.

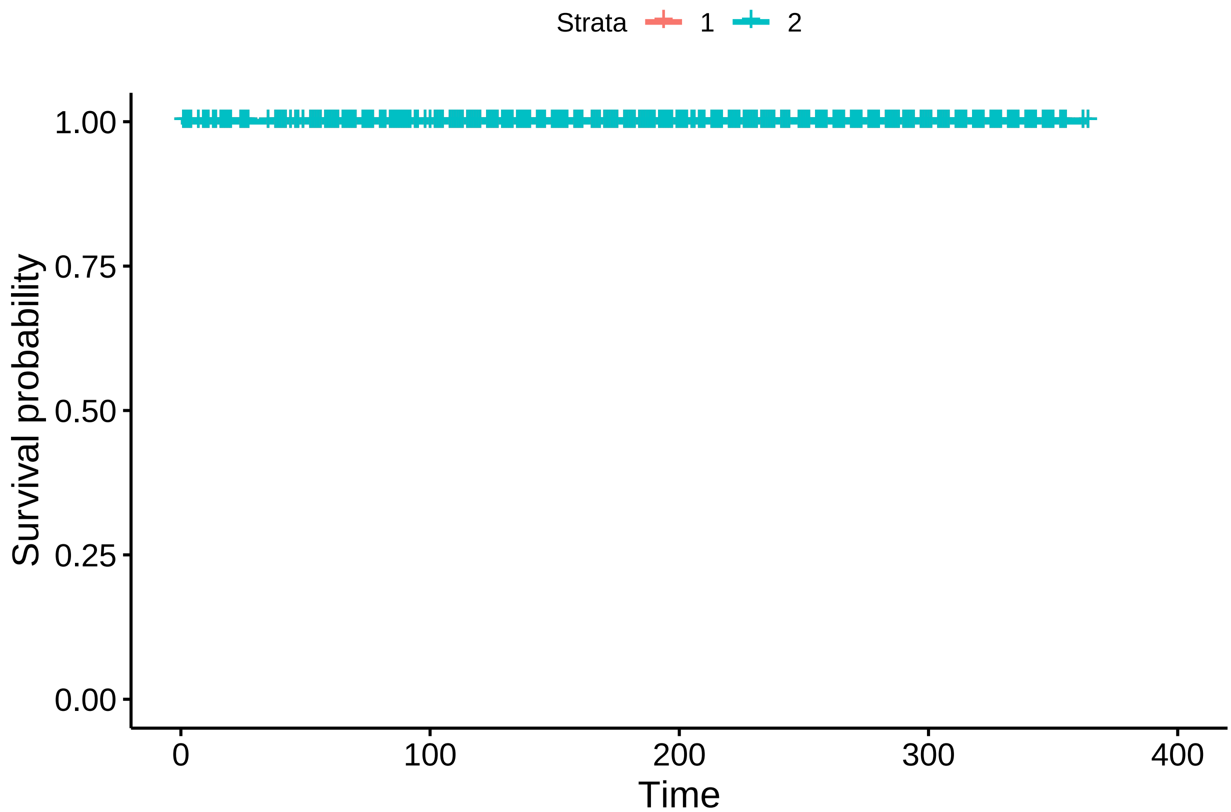
The coefficients of both variables are very low, therefore, we can conclude that the less diseased and younger someone is the lower their risk and higher their survival rate.

Part (f)

```
#new_data <- data.frame(patient = c(1,2), treatment = c(1,0), CD4 = c(86,20), sex = c(2,2), IV = c(1,2)
new_data <- data.frame(V1 = c(1,2), V6 = c(1,0), V8 = c(86,20), V9 = c(2,2), V11 = c(1,2), V12 = c(0,1)
head(new_data)
```

```
##   V1 V6 V8 V9 V11 V12 V13 V15 V16
## 1  1  1 86  2   1   0  90  30  38
## 2  2  0 20  2   2   1  70 250  25
```

```
fit2 <- survfit(CPH, newdata = new_data)
ggsurvplot(fit2, data = new_data, conf.int = FALSE)
```




```
#CPHnew <- coxph(formula = Surv(V2, V5) ~ V6 + V8 + V9 + V11 + V12 + V13 + V15 + V16, data = new_data)

#plot(survfit(CPHnew), main = 'CPH new patients', xlab = 'Days',  ylab = 'Survival')
```

Again, followed lab to get a result inconsistent with what is being asked for homework?