Variable Selection

Cindy J. Pang

2025-03-18

```
rm(list=ls())
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(survival)
library(survMisc)
```

```
data <- read.csv("../data/survDataCleaned.csv")
```

Which race variable should we use?

```
data.raceCleaned <- data %>% select(-race_collapsed)
data.raceCollapsed <- data %>% select(-race_cleaned)
```

Variable Selection.

Fit Cox-PH models for both the `data.raceCleaned` and `data.raceCollapsed`.

```
data.raceCleaned$race_cleaned <- relevel(factor(data.raceCleaned$race_cleaned), ref = "white")
data.raceCleaned$FIGO <- relevel(factor(data.raceCleaned$FIGO), ref = "Stage I")

fullCoxMod <- coxph(Surv(time, delta) ~ size.intermediate + factor(race_cleaned) + factor(FIGO)+age_at_

fullCoxMod
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
##     factor(FIGO) + age_at_diagnosis, data = data.raceCleaned)
##
##                                           coef exp(coef)  se(coef)      z
## size.intermediate                     0.151974  1.164130  0.147632  1.029
## factor(race_cleaned)asian            -0.914762  0.400612  0.413712 -2.211
## factor(race_cleaned)black             0.280853  1.324259  0.221897  1.266
## factor(race_cleaned)hispanic          0.093532  1.098046  0.384043  0.244
## factor(race_cleaned)unreported/other -0.896446  0.408017  0.383582 -2.337
## factor(FIGO)Stage II                  0.578492  1.783347  0.660023  0.876
## factor(FIGO)Stage III                 1.546288  4.694015  0.581007  2.661
## factor(FIGO)Stage IV                  2.019321  7.533212  0.591642  3.413
## age_at_diagnosis                      0.021200  1.021426  0.004712  4.499
##                                             p
## size.intermediate                    0.303286
## factor(race_cleaned)asian            0.027028
## factor(race_cleaned)black            0.205624
## factor(race_cleaned)hispanic         0.807583
## factor(race_cleaned)unreported/other 0.019437
## factor(FIGO)Stage II                 0.380773
## factor(FIGO)Stage III                0.007782
## factor(FIGO)Stage IV                 0.000642
## age_at_diagnosis                     6.83e-06
##
## Likelihood ratio test=78.46  on 9 df, p=3.264e-13
## n= 598, number of events= 359
```

```
data.raceCollapsed$race_cleaned <- relevel(factor(data.raceCollapsed$race_collapsed), ref = "white")
data.raceCollapsed$FIGO <- relevel(factor(data.raceCollapsed$FIGO), ref = "Stage I")

fullCoxMod.r <- coxph(Surv(time, delta) ~ size.intermediate + factor(race_cleaned) + factor(FIGO)+age_a

fullCoxMod.r
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
##     factor(FIGO) + age_at_diagnosis, data = data.raceCollapsed)
##
##                                           coef exp(coef)  se(coef)      z
## size.intermediate                     0.153985  1.166473  0.147402  1.045
## factor(race_cleaned)asian            -0.913906  0.400955  0.413715 -2.209
## factor(race_cleaned)black/hisp.       0.232315  1.261517  0.195233  1.190
## factor(race_cleaned)other/unreported -0.896253  0.408096  0.383577 -2.337
## factor(FIGO)Stage II                  0.583044  1.791483  0.659915  0.884
## factor(FIGO)Stage III                 1.546116  4.693204  0.581015  2.661
## factor(FIGO)Stage IV                  2.017950  7.522891  0.591655  3.411
## age_at_diagnosis                      0.021367  1.021597  0.004697  4.549
##                                             p
## size.intermediate                    0.296180
## factor(race_cleaned)asian            0.027173
## factor(race_cleaned)black/hisp.      0.234071
## factor(race_cleaned)other/unreported 0.019462
## factor(FIGO)Stage II                 0.376959
## factor(FIGO)Stage III                0.007790
## factor(FIGO)Stage IV                 0.000648
## age_at_diagnosis                     5.39e-06
##
## Likelihood ratio test=78.27  on 8 df, p=1.089e-13
## n= 598, number of events= 359
```

Above is proof that collapsing the race category doesn’t improve or make inference any better. I will use the `data.raceCleaned` variable going forward.

Stepwise AIC Selection.

```r
# import package
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
best.model.aic <- stepAIC(fullCoxMod, direction="both",
                          k=2, trace=1)
```

```
## Start:  AIC=4247.61
## Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
##     factor(FIGO) + age_at_diagnosis
##
##                         Df    AIC
## - size.intermediate     1 4246.6
## <none>                    4247.6
## - factor(race_cleaned)  4 4255.3
## - age_at_diagnosis      1 4265.7
## - factor(FIGO)          3 4279.7
##
## Step:  AIC=4246.65
## Surv(time, delta) ~ factor(race_cleaned) + factor(FIGO) + age_at_diagnosis
##
##                         Df    AIC
## <none>                    4246.6
## + size.intermediate     1 4247.6
## - factor(race_cleaned)  4 4254.9
## - age_at_diagnosis      1 4265.7
## - factor(FIGO)          3 4279.5
```

```r
# Display the best model obtained by AIC
best.model.aic
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ factor(race_cleaned) + factor(FIGO) +
##     age_at_diagnosis, data = data.raceCleaned)
##
##                                           coef exp(coef)  se(coef)      z
## factor(race_cleaned)asian            -0.928315  0.395219  0.413392 -2.246
## factor(race_cleaned)black             0.287141  1.332613  0.221873  1.294
## factor(race_cleaned)hispanic          0.085670  1.089447  0.383921  0.223
## factor(race_cleaned)unreported/other -0.907432  0.403559  0.383402 -2.367
## factor(FIGO)Stage II                  0.579955  1.785958  0.660003  0.879
## factor(FIGO)Stage III                 1.544113  4.683814  0.580996  2.658
## factor(FIGO)Stage IV                  2.035096  7.652983  0.591420  3.441
## age_at_diagnosis                      0.021626  1.021861  0.004692  4.609
##                                             p
## factor(race_cleaned)asian            0.024729
## factor(race_cleaned)black            0.195606
## factor(race_cleaned)hispanic         0.823422
## factor(race_cleaned)unreported/other 0.017943
## factor(FIGO)Stage II                 0.379556
## factor(FIGO)Stage III                0.007868
## factor(FIGO)Stage IV                 0.000579
## age_at_diagnosis                     4.04e-06
##
## Likelihood ratio test=77.43  on 8 df, p=1.608e-13
## n= 598, number of events= 359
```

Ok, AIC stepwise selection only selected `race_cleaned`, `FIGO` and `age_at_diagnosis`.

Stepwise BIC Selection.

```
n <- nrow(data) # number of rows
best.model.bic <- stepAIC(fullCoxMod, direction="both",
                          k=log(n), trace=1)
```

```
## Start:  AIC=4287.15
## Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
##     factor(FIGO) + age_at_diagnosis
##
##                        Df    AIC
## - factor(race_cleaned)  4 4277.3
## - size.intermediate     1 4281.8
## <none>                    4287.2
## - age_at_diagnosis      1 4300.8
## - factor(FIGO)          3 4306.0
##
## Step:  AIC=4277.28
## Surv(time, delta) ~ size.intermediate + factor(FIGO) + age_at_diagnosis
##
##                        Df    AIC
## - size.intermediate     1 4272.5
## <none>                    4277.3
## + factor(race_cleaned)  4 4287.2
## - age_at_diagnosis      1 4293.9
## - factor(FIGO)          3 4298.4
##
## Step:  AIC=4272.49
## Surv(time, delta) ~ factor(FIGO) + age_at_diagnosis
##
##                        Df    AIC
## <none>                    4272.5
## + size.intermediate     1 4277.3
## + factor(race_cleaned)  4 4281.8
## - age_at_diagnosis      1 4290.5
## - factor(FIGO)          3 4294.6
```

```
best.model.bic
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ factor(FIGO) + age_at_diagnosis,
##     data = data.raceCleaned)
##
##                            coef exp(coef) se(coef)     z        p
## factor(FIGO)Stage II  0.543453  1.721942 0.658470 0.825 0.409186
## factor(FIGO)Stage III 1.534167  4.637460 0.580808 2.641 0.008256
## factor(FIGO)Stage IV  2.050814  7.774229 0.591210 3.469 0.000523
## age_at_diagnosis      0.022862  1.023125 0.004608 4.961    7e-07
##
## Likelihood ratio test=61.16  on 4 df, p=1.655e-12
## n= 598, number of events= 359
```

BIC selected `FIGO` and `age_at_diagonosis` to obtain the best model.

AFT Models.

Exponential AFT.

```
full.expAFT <- survreg(Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
    factor(FIGO) + age_at_diagnosis, data = data.raceCleaned, dist = "exponential")

expAFT.aic <- stepAIC(full.expAFT, direction = "both", k=2, trace=1)
```

```
## Start:  AIC=6673.96
## Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
##     factor(FIGO) + age_at_diagnosis
##
##                         Df    AIC
## <none>                    6674.0
## - size.intermediate     1 6674.9
## - factor(race_cleaned)  4 6687.1
## - age_at_diagnosis      1 6692.7
## - factor(FIGO)          3 6717.2
```

```
expAFT.aic
```

```
## Call:
## survreg(formula = Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
##     factor(FIGO) + age_at_diagnosis, data = data.raceCleaned,
##     dist = "exponential")
##
## Coefficients:
##                          (Intercept)                    size.intermediate
##                           11.5786835                           -0.2600141
##            factor(race_cleaned)asian            factor(race_cleaned)black
##                            0.9821891                           -0.3971982
##         factor(race_cleaned)hispanic factor(race_cleaned)unreported/other
##                           -0.1746070                            1.0148055
##                 factor(FIGO)Stage II                factor(FIGO)Stage III
##                           -0.6588743                           -1.7113719
##                 factor(FIGO)Stage IV                     age_at_diagnosis
##                           -2.2627941                           -0.0213537
##
## Scale fixed at 1
##
## Loglik(model)= -3327   Loglik(intercept only)= -3377.4
##  Chisq= 100.91 on 9 degrees of freedom, p= <2e-16
## n= 598
```

```r
expAFT.bic <- stepAIC(full.expAFT, direction = "both", k=log(n), trace=1)
```

```
## Start:  AIC=6717.9
## Surv(time, delta) ~ size.intermediate + factor(race_cleaned) + 
##     factor(FIGO) + age_at_diagnosis
## 
##                         Df    AIC
## - factor(race_cleaned)  4 6713.5
## - size.intermediate     1 6714.5
## <none>                    6717.9
## - age_at_diagnosis      1 6732.2
## - factor(FIGO)          3 6748.0
## 
## Step:  AIC=6713.49
## Surv(time, delta) ~ size.intermediate + factor(FIGO) + age_at_diagnosis
## 
##                         Df    AIC
## - size.intermediate     1 6711.3
## <none>                    6713.5
## + factor(race_cleaned)  4 6717.9
## - age_at_diagnosis      1 6731.0
## - factor(FIGO)          3 6746.6
## 
## Step:  AIC=6711.3
## Surv(time, delta) ~ factor(FIGO) + age_at_diagnosis
## 
##                         Df    AIC
## <none>                    6711.3
## + size.intermediate     1 6713.5
## + factor(race_cleaned)  4 6714.5
## - age_at_diagnosis      1 6731.1
## - factor(FIGO)          3 6746.3
```

```r
expAFT.bic
```

```
## Call:
## survreg(formula = Surv(time, delta) ~ factor(FIGO) + age_at_diagnosis, 
##     data = data.raceCleaned, dist = "exponential")
## 
## Coefficients:
##           (Intercept)  factor(FIGO)Stage II factor(FIGO)Stage III 
##           11.52387703           -0.64591491           -1.71559843 
##  factor(FIGO)Stage IV      age_at_diagnosis 
##           -2.32689756           -0.02338033 
## 
## Scale fixed at 1 
## 
## Loglik(model)= -3339.7   Loglik(intercept only)= -3377.4
##  Chisq= 75.54 on 4 degrees of freedom, p= 1.53e-15 
## n= 598
```

Weibull AFT

```r
full.weibullAFT <- survreg(Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
    factor(FIGO) + age_at_diagnosis, data = data.raceCleaned, dist = "weibull")

weibullAFT.aic <- stepAIC(full.weibullAFT, direction = "both", k=2, trace=1)
```

```
## Start:  AIC=6640.12
## Surv(time, delta) ~ size.intermediate + factor(race_cleaned) + 
##     factor(FIGO) + age_at_diagnosis
## 
##                         Df    AIC
## <none>                     6640.1
## - size.intermediate     1 6640.4
## - factor(race_cleaned)  4 6650.7
## - age_at_diagnosis      1 6656.5
## - factor(FIGO)          3 6676.7
```

```r
weibullAFT.aic
```

```
## Call:
## survreg(formula = Surv(time, delta) ~ size.intermediate + factor(race_cleaned) + 
##     factor(FIGO) + age_at_diagnosis, data = data.raceCleaned, 
##     dist = "weibull")
## 
## Coefficients:
##                           (Intercept)                    size.intermediate 
##                           12.36301594                          -0.28985830 
##             factor(race_cleaned)asian            factor(race_cleaned)black 
##                            1.22733827                          -0.44097742 
##          factor(race_cleaned)hispanic factor(race_cleaned)unreported/other 
##                           -0.20784887                           1.24874844 
##                  factor(FIGO)Stage II                factor(FIGO)Stage III 
##                           -0.82430720                          -2.11463119 
##                  factor(FIGO)Stage IV                     age_at_diagnosis 
##                           -2.76756691                          -0.02613169 
## 
## Scale= 1.294644 
## 
## Loglik(model)= -3309.1   Loglik(intercept only)= -3352.5
##  Chisq= 86.94 on 9 degrees of freedom, p= 6.68e-15 
## n= 598
```

```r
weibullAFT.bic <- stepAIC(full.weibullAFT, direction = "both", k=log(n), trace=1)
```

```
## Start:  AIC=6688.45
## Surv(time, delta) ~ size.intermediate + factor(race_cleaned) + 
##     factor(FIGO) + age_at_diagnosis
## 
##                         Df    AIC
## - factor(race_cleaned)  4 6681.4
## - size.intermediate     1 6684.3
## <none>                     6688.4
## - age_at_diagnosis      1 6700.4
## - factor(FIGO)          3 6711.8
## 
## Step:  AIC=6681.41
## Surv(time, delta) ~ size.intermediate + factor(FIGO) + age_at_diagnosis
##
##                         Df    AIC
## - size.intermediate      1 6678.2
## <none>                     6681.4
## + factor(race_cleaned)   4 6688.4
## - age_at_diagnosis       1 6696.2
## - factor(FIGO)           3 6707.3
##
## Step:  AIC=6678.2
## Surv(time, delta) ~ factor(FIGO) + age_at_diagnosis
##
##                         Df    AIC
## <none>                     6678.2
## + size.intermediate      1 6681.4
## + factor(race_cleaned)   4 6684.3
## - age_at_diagnosis       1 6694.8
## - factor(FIGO)           3 6705.6
```

```
weibullAFT.bic
```

```
## Call:
## survreg(formula = Surv(time, delta) ~ factor(FIGO) + age_at_diagnosis,
##     data = data.raceCleaned, dist = "weibull")
##
## Coefficients:
##           (Intercept)  factor(FIGO)Stage II factor(FIGO)Stage III
##           12.35580783           -0.80343728           -2.13537950
##  factor(FIGO)Stage IV      age_at_diagnosis
##           -2.86714404           -0.02884478
##
## Scale= 1.311245
##
## Loglik(model)= -3319.9   Loglik(intercept only)= -3352.5
##  Chisq= 65.22 on 4 degrees of freedom, p= 2.31e-13
## n= 598
```

Log-Logistic AFT.

```
full.loglogisticAFT <- survreg(Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
    factor(FIGO) + age_at_diagnosis, data = data.raceCleaned, dist = "loglogistic")

loglogisticAFT.aic <- stepAIC(full.loglogisticAFT, direction = "both", k=2, trace=1)
```

```
## Start:  AIC=6595.7
## Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
##     factor(FIGO) + age_at_diagnosis
##
##                         Df    AIC
## - size.intermediate      1 6594.6
## <none>                     6595.7
## - factor(race_cleaned)   4 6605.0
## - age_at_diagnosis       1 6618.7
## - factor(FIGO)           3 6630.9
##
```

```
## Step:  AIC=6594.55
## Surv(time, delta) ~ factor(race_cleaned) + factor(FIGO) + age_at_diagnosis
##
##                         Df    AIC
## <none>                     6594.6
## + size.intermediate     1 6595.7
## - factor(race_cleaned)  4 6604.2
## - age_at_diagnosis      1 6618.3
## - factor(FIGO)          3 6630.5
```

```
loglogisticAFT.aic
```

```
## Call:
## survreg(formula = Surv(time, delta) ~ factor(race_cleaned) + 
##     factor(FIGO) + age_at_diagnosis, data = data.raceCleaned, 
##     dist = "loglogistic")
##
## Coefficients:
##                          (Intercept)            factor(race_cleaned)asian 
##                          11.75168203                           1.24102582 
##            factor(race_cleaned)black         factor(race_cleaned)hispanic 
##                          -0.48585138                          -0.17884472 
## factor(race_cleaned)unreported/other                 factor(FIGO)Stage II 
##                           1.09502164                          -0.72098231 
##                factor(FIGO)Stage III                 factor(FIGO)Stage IV 
##                          -1.96859618                          -2.69683380 
##                     age_at_diagnosis 
##                          -0.03253153 
##
## Scale= 0.9524551 
##
## Loglik(model)= -3287.3   Loglik(intercept only)= -3330.7
##  Chisq= 86.94 on 8 degrees of freedom, p= 1.94e-15 
## n= 598
```

```
loglogisticAFT.bic <- stepAIC(full.loglogisticAFT, direction = "both", k=log(n), trace=1)
```

```
## Start:  AIC=6644.03
## Surv(time, delta) ~ size.intermediate + factor(race_cleaned) + 
##     factor(FIGO) + age_at_diagnosis
##
##                         Df    AIC
## - factor(race_cleaned)  4 6635.7
## - size.intermediate     1 6638.5
## <none>                     6644.0
## - age_at_diagnosis      1 6662.6
## - factor(FIGO)          3 6666.1
##
## Step:  AIC=6635.75
## Surv(time, delta) ~ size.intermediate + factor(FIGO) + age_at_diagnosis
##
##                         Df    AIC
## - size.intermediate     1 6630.6
## <none>                     6635.7
## + factor(race_cleaned)  4 6644.0
## - age_at_diagnosis      1 6657.3
## - factor(FIGO)          3 6659.9
##
## Step:  AIC=6630.58
## Surv(time, delta) ~ factor(FIGO) + age_at_diagnosis
##
##                        Df    AIC
## <none>                    6630.6
## + size.intermediate     1 6635.7
## + factor(race_cleaned)  4 6638.5
## - age_at_diagnosis      1 6653.1
## - factor(FIGO)          3 6655.7
```

```
loglogisticAFT.bic
```

```
## Call:
## survreg(formula = Surv(time, delta) ~ factor(FIGO) + age_at_diagnosis,
##     data = data.raceCleaned, dist = "loglogistic")
##
## Coefficients:
##          (Intercept)  factor(FIGO)Stage II factor(FIGO)Stage III
##          11.93419596           -0.69099201           -1.98625823
##  factor(FIGO)Stage IV      age_at_diagnosis
##          -2.75224858           -0.03446027
##
## Scale= 0.9666854
##
## Loglik(model)= -3296.1   Loglik(intercept only)= -3330.7
##  Chisq= 69.27 on 4 degrees of freedom, p= 3.24e-14
## n= 598
```

Log-Normal AFT

```
full.lognormalAFT <- survreg(Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
    factor(FIGO) + age_at_diagnosis, data = data.raceCleaned, dist = "lognormal")

lognormalAFT.aic <- stepAIC(full.lognormalAFT, direction = "both", k=2, trace=1)
```

```
## Start:  AIC=6604.08
## Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
##     factor(FIGO) + age_at_diagnosis
##
##                        Df    AIC
## - size.intermediate     1 6602.9
## <none>                    6604.1
## - factor(race_cleaned)  4 6611.1
## - age_at_diagnosis      1 6629.1
## - factor(FIGO)          3 6638.9
##
## Step:  AIC=6602.87
## Surv(time, delta) ~ factor(race_cleaned) + factor(FIGO) + age_at_diagnosis
##
##                        Df    AIC
## <none>                    6602.9
## + size.intermediate     1 6604.1
## - factor(race_cleaned)  4 6610.4
## - age_at_diagnosis      1 6628.8
## - factor(FIGO)          3 6638.3
```

```
lognormalAFT.aic
```

```
## Call:
## survreg(formula = Surv(time, delta) ~ factor(race_cleaned) +
##     factor(FIGO) + age_at_diagnosis, data = data.raceCleaned,
##     dist = "lognormal")
##
## Coefficients:
##                          (Intercept)            factor(race_cleaned)asian
##                          11.60283189                           1.32670368
##            factor(race_cleaned)black         factor(race_cleaned)hispanic
##                          -0.60372256                          -0.21390418
## factor(race_cleaned)unreported/other                 factor(FIGO)Stage II
##                           0.71779553                          -0.27403185
##                factor(FIGO)Stage III                 factor(FIGO)Stage IV
##                          -1.53718723                          -2.39908988
##                     age_at_diagnosis
##                          -0.03567528
##
## Scale= 1.698155
##
## Loglik(model)= -3291.4   Loglik(intercept only)= -3333.2
##  Chisq= 83.54 on 8 degrees of freedom, p= 9.43e-15
## n= 598
```

```
lognormalAFT.bic <- stepAIC(full.lognormalAFT, direction = "both", k=log(n), trace=1)
```

```
## Start:  AIC=6652.41
## Surv(time, delta) ~ size.intermediate + factor(race_cleaned) +
##     factor(FIGO) + age_at_diagnosis
##
##                          Df    AIC
## - factor(race_cleaned)  4 6641.9
## - size.intermediate     1 6646.8
## <none>                    6652.4
## - age_at_diagnosis      1 6673.0
## - factor(FIGO)          3 6674.0
##
## Step:  AIC=6641.89
## Surv(time, delta) ~ size.intermediate + factor(FIGO) + age_at_diagnosis
##
##                          Df    AIC
## - size.intermediate     1 6636.8
## <none>                    6641.9
## + factor(race_cleaned)  4 6652.4
## - factor(FIGO)          3 6663.8
## - age_at_diagnosis      1 6664.4
##
## Step:  AIC=6636.78
## Surv(time, delta) ~ factor(FIGO) + age_at_diagnosis
##
```

```
##                         Df    AIC
## <none>                     6636.8
## + size.intermediate      1 6641.9
## + factor(race_cleaned)  4 6646.8
## - factor(FIGO)           3 6659.4
## - age_at_diagnosis      1 6660.5
```

```r
lognormalAFT.bic
```

```
## Call:
## survreg(formula = Surv(time, delta) ~ factor(FIGO) + age_at_diagnosis,
##     data = data.raceCleaned, dist = "lognormal")
##
## Coefficients:
##          (Intercept)  factor(FIGO)Stage II factor(FIGO)Stage III
##           11.6248024            -0.1457258            -1.4340866
##  factor(FIGO)Stage IV      age_at_diagnosis
##           -2.3211559            -0.0371437
##
## Scale= 1.716158
##
## Loglik(model)= -3299.2   Loglik(intercept only)= -3333.2
##  Chisq= 68 on 4 degrees of freedom, p= 6.01e-14
## n= 598
```

Conclusions.

For both the AIC and BIC Stepwise selection models, `FIGO` and `age_at_diagnosis` were selected. In both models, the Biopsy size variable `size.intermediate` was not considered significant. Going forward, it would be prudent to **use the BIC chosen model** since there is a larger penalty which adjusts for the large sample size (n=598).

Model Diagnostics.

Cox-Snell Residual Plot.

```r
plotCoxSnellCPH<- function(survFit, delta, fitType){

  # get Cox-Snell Residual based on Martingale Residuals
  mg.residual <- resid(survFit, type = "martingale")

  cs.residual <- delta - mg.residual

  # Graphical Plot
  fit.cs <- survfit(Surv(cs.residual, delta) ~ 1) # get KM estimates
  H.cs <- cumsum(fit.cs$n.event/fit.cs$n.risk)

  plot(fit.cs$time, H.cs, type='s', col='blue',
       main = paste0('Cox-PH - ', fitType),
       xlab = 'Residual', ylab = 'Nelson-Aalen Cum. Hazard')
  #Note here that 'time' is the value of the Cox-Snell residual
  abline(0, 1, col='red',  lty = 2)
}
```

```
plotCSExpAFT <- function(survFit, data, fitType){
  sigma <- survFit$scale
  eta   <- -survFit$linear.predictors/sigma

  r.exp <- data$time * exp(eta)

  fit   <- survfit(Surv(r.exp, data$delta) ~ 1)
  H.exp <- cumsum(fit$n.event / fit$n.risk)

  plot(H.exp ~ fit$time, type = 'l', main = paste0('Exponential AFT - ', fitType),
       ylab = 'Estimated Cumulative Hazard', xlab = 'Cox-Snell Residual')
  abline(0, 1, col='red',  lty=2)
}
```

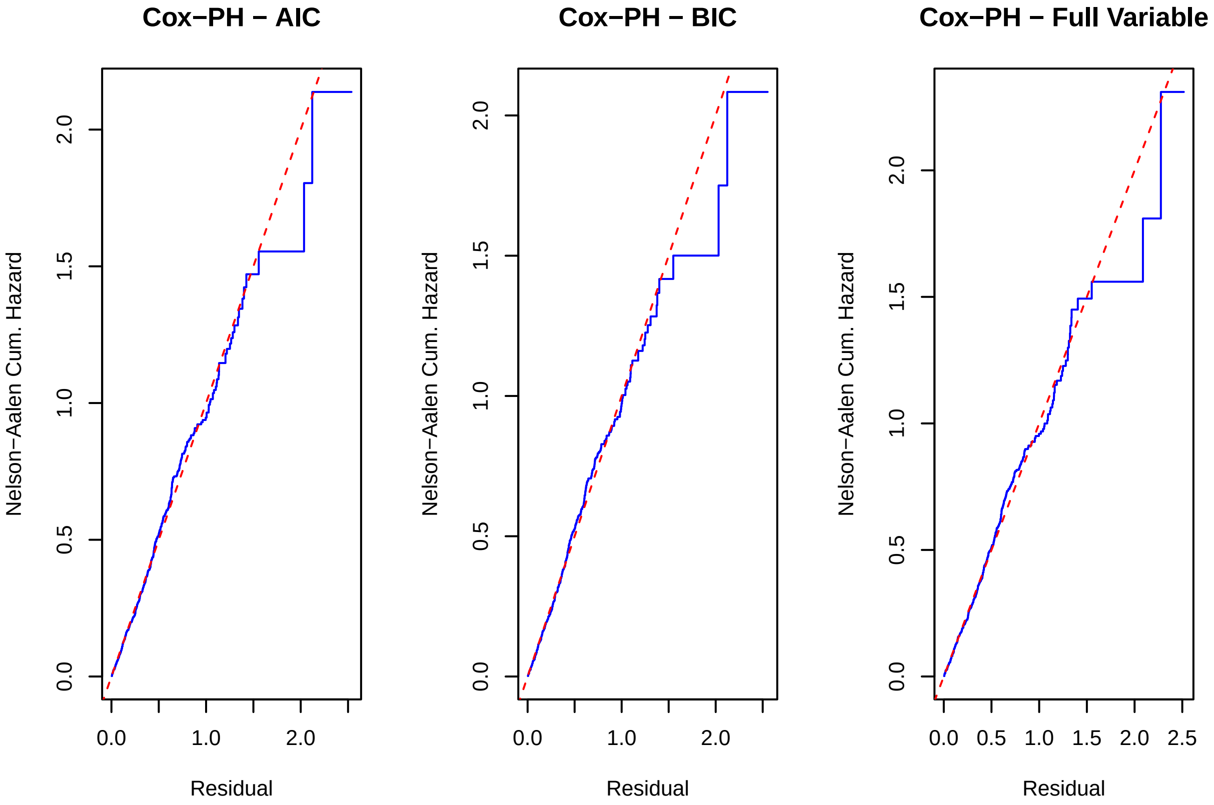
```
plotCSWeibullAFT <- function(survFit, data, fitType){
  sigma  <- survFit$scale
  alpha  <- 1 / sigma
  eta    <- -survFit$linear.predictors / sigma

  r.wb <- data$time^alpha * exp(eta)

  fit   <- survfit(Surv(r.wb, data$delta) ~ 1)
  H.wb  <- cumsum(fit$n.event/fit$n.risk)

  plot(H.wb ~ fit$time, type = 'l', main = paste0('Weibull AFT - ', fitType),
       ylab = 'Estimated Cumulative Hazard', xlab = 'Cox-Snell Residual')
  abline(0, 1, col='red',  lty=2)
}
```

```
plotCSLogLogisticAFT <- function(survFit, data, fitType){
  sigma  <- data$scale
  alpha  <- 1 / sigma
  eta    <- -data$linear.predictors / sigma

  r.ll  <- -log(1/(1 + data$time^alpha*exp(eta)))

  fit   <- survfit(Surv(r.ll, data$delta) ~ 1)
  H.ll  <- cumsum(fit$n.event / fit$n.risk)

  plot(H.ll ~ fit$time, type = 'l', main = paste0('Log-Logistic AFT- ', fitType),
       ylab = 'Estimated Cumulative Hazard', xlab = 'Cox-Snell Residual')
  abline(0, 1, col='red',  lty=2)
}
```

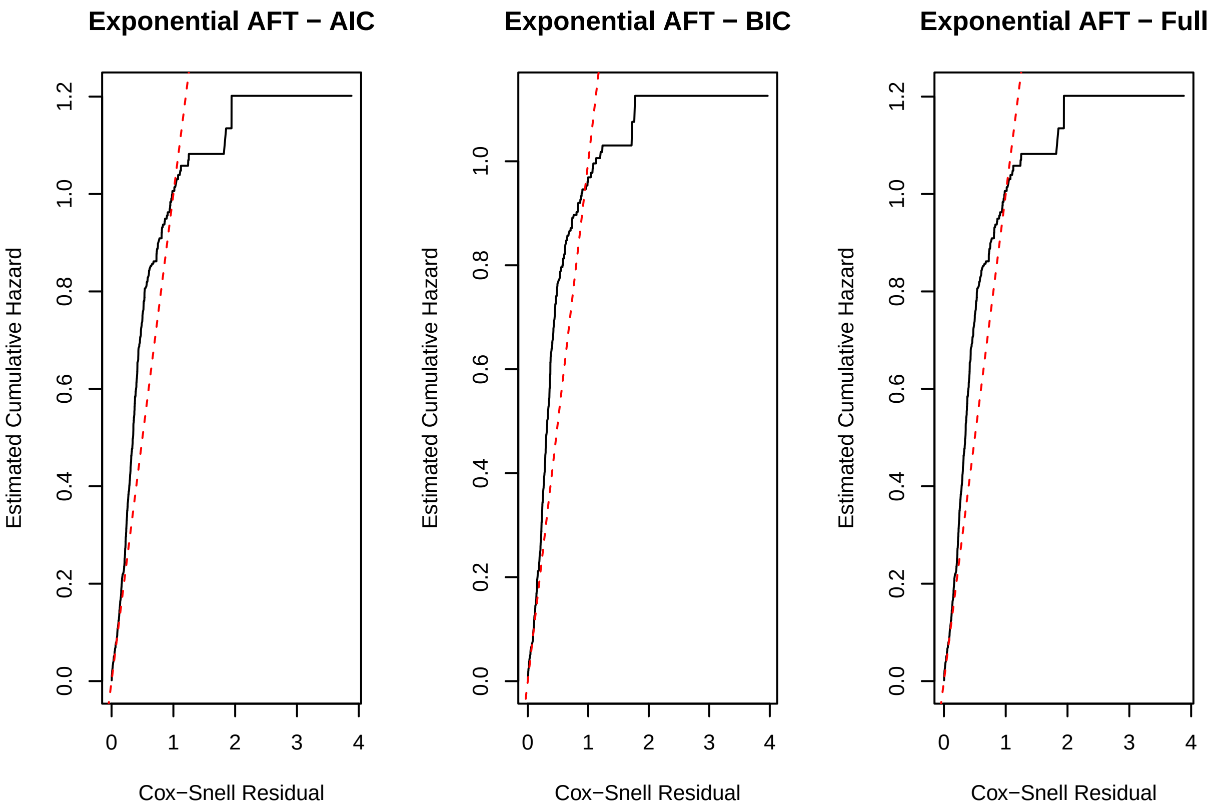
```
plotCSLogNormalAFT <- function(survFit, data, fitType){
  eta    <- -survFit$linear.predictors / survFit$scale
  r.ln   <- -log(1 - pnorm((log(data$time) - survFit$linear.predictors) / survFit$scale))

  fit   <- survfit(Surv(r.ln, data$delta) ~ 1)
  H.ln  <- cumsum(fit$n.event/fit$n.risk)

  plot(H.ln ~ fit$time, type = 'l', main = paste0('Log-Normal AFT - ', fitType),
       ylab = 'Estimated Cumulative Hazard', xlab = 'Cox-Snell Residual')
  abline(0, 1, col='red',  lty = 2)
}
```

```
bic.modfit <- coxph(Surv(time, delta)~factor(FIGO)+age_at_diagnosis, data = data.raceCleaned, method =
aic.modfit <- coxph(Surv(time,delta)~factor(race_cleaned)+factor(FIGO)+age_at_diagnosis, data = data.ra
```

```
par(mfrow = c(1,3))
plotCoxSnellCPH(aic.modfit, data.raceCleaned$delta, "AIC")
plotCoxSnellCPH(bic.modfit, data.raceCleaned$delta, "BIC")
plotCoxSnellCPH(fullCoxMod, data.raceCleaned$delta, "Full Variable")
```



```
par(mfrow = c(1,3))
plotCSExpAFT(expAFT.aic, data.raceCleaned, "AIC")
plotCSExpAFT(expAFT.bic, data.raceCleaned, "BIC")
plotCSExpAFT(full.expAFT, data.raceCleaned, "Full")
```



Ok both models fit similarly, which isn't helpful.