Homework3

Yuki Joyama

```
# import MI data
df = haven::read_dta("./data/actg320.dta")
```

1. Model Selection

Endpoint: time to AIDS progression or death (time)
Primary variable of interest: treatment with IDV or no IDV

I will use Collett's model selection approach to decide the final model.
First, let's fit univariate models for each covariate and identify the predictors significant at $\alpha = 0.20$.

```
covariates = c("trt", "hemophil", "hemophil", "karnof", "cd4", "priorzdv", "age",
               "female", "sqrtcd4", "cd4_50", "cd4cat", "cd4_under50", "cd450_100",
               "cd4_squared", "ivdu", "black", "hispanic", "karnof100")
univariate = list()

# function for univariate model
for (cov in covariates) {
  formula <- as.formula(paste("Surv(time, censor) ~", cov))
  model <- coxph(formula, data = df)
  summary <- summary(model)
  p_value <- summary$coefficients[1, "Pr(>|z|)"]
  univariate[[cov]] <- p_value
}

# filter covariates with p-value <= 0.20
sig_cov <- names(univariate[univariate <= 0.20])
print(sig_cov)
##  [1] "trt"         "karnof"      "cd4"         "age"         "sqrtcd4"
##  [6] "cd4_50"      "cd4cat"      "cd4_under50" "cd4_squared" "karnof100"
```

Listed covariates meet p-value < 0.20. Now, I will move on to evaluate multivariate model with all significant univariate predictors and use backward selection to eliminate non-significant variables at level 0.10.

```
# multivariate Cox model using significant univariate predictors
mult_formula <- as.formula(paste("Surv(time, censor) ~", paste(sig_cov, collapse = " + ")))
multivariate <- coxph(mult_formula, data = df)

# perform backward selection with p-value threshold of 0.10
bkwd <- step(multivariate, direction = "backward", k = qchisq(0.10, 1, lower.tail=FALSE)) # check param
## Start:  AIC=1237.9
## Surv(time, censor) ~ trt + karnof + cd4 + age + sqrtcd4 + cd4_50 +
```

```
##     cd4cat + cd4_under50 + cd4_squared + karnof100
##
##
## Step:  AIC=1237.9
## Surv(time, censor) ~ trt + karnof + cd4 + age + sqrtcd4 + cd4cat +
##     cd4_under50 + cd4_squared + karnof100
##
##               Df    AIC
## - karnof100    1 1235.3
## - cd4cat       1 1235.4
## - cd4_under50  1 1235.9
## - cd4_squared  1 1237.5
## - sqrtcd4      1 1237.7
## <none>           1237.9
## - age          1 1238.4
## - cd4          1 1240.0
## - trt          1 1245.0
## - karnof       1 1246.7
##
## Step:  AIC=1235.31
## Surv(time, censor) ~ trt + karnof + cd4 + age + sqrtcd4 + cd4cat +
##     cd4_under50 + cd4_squared
##
##               Df    AIC
## - cd4cat       1 1232.8
## - cd4_under50  1 1233.3
## - cd4_squared  1 1234.8
## - sqrtcd4      1 1235.0
## <none>           1235.3
## - age          1 1235.9
## - cd4          1 1237.4
## - trt          1 1242.5
## - karnof       1 1252.8
##
## Step:  AIC=1232.83
## Surv(time, censor) ~ trt + karnof + cd4 + age + sqrtcd4 + cd4_under50 +
##     cd4_squared
##
##               Df    AIC
## - cd4_squared  1 1232.2
## - sqrtcd4      1 1232.4
## <none>           1232.8
## - cd4_under50  1 1233.0
## - age          1 1233.5
## - cd4          1 1235.0
## - trt          1 1240.0
## - karnof       1 1250.3
##
## Step:  AIC=1232.17
## Surv(time, censor) ~ trt + karnof + cd4 + age + sqrtcd4 + cd4_under50
##
##               Df    AIC
## - sqrtcd4      1 1229.8
```

```
## - cd4_under50  1 1230.7
## <none>           1232.2
## - age          1 1232.8
## - cd4          1 1237.3
## - trt          1 1239.4
## - karnof       1 1249.9
##
## Step:  AIC=1229.83
## Surv(time, censor) ~ trt + karnof + cd4 + age + cd4_under50
##
##               Df    AIC
## - cd4_under50  1 1228.6
## <none>           1229.8
## - age          1 1230.5
## - trt          1 1237.0
## - karnof       1 1247.2
## - cd4          1 1249.4
##
## Step:  AIC=1228.63
## Surv(time, censor) ~ trt + karnof + cd4 + age
##
##          Df    AIC
## <none>      1228.6
## - age     1 1229.5
## - trt     1 1235.9
## - karnof  1 1245.8
## - cd4     1 1274.8

summary(bkwd)
## Call:
## coxph(formula = Surv(time, censor) ~ trt + karnof + cd4 + age,
##     data = df)
##
##   n= 1151, number of events= 96
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## trt    -0.661203  0.516230  0.215139 -3.073  0.00212 **
## karnof -0.053485  0.947920  0.011821 -4.525 6.05e-06 ***
## cd4    -0.014554  0.985552  0.002512 -5.793 6.92e-09 ***
## age     0.021879  1.022120  0.011349  1.928  0.05387 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## trt       0.5162     1.9371    0.3386    0.7870
## karnof    0.9479     1.0549    0.9262    0.9701
## cd4       0.9856     1.0147    0.9807    0.9904
## age       1.0221     0.9784    0.9996    1.0451
##
## Concordance= 0.781  (se = 0.023 )
## Likelihood ratio test= 99.07  on 4 df,   p=<2e-16
## Wald test            = 81.73  on 4 df,   p=<2e-16
## Score (logrank) test = 91.87  on 4 df,   p=<2e-16
```

Next, I will perform forward selection to consider each of the non-significant variable with significance level of 0.10.

```
sig_cov <- names(coef(bkwd))  # extract significant covariates from backward model

# remove variables not present in the backward model from the lower scope
fwd_formula <- as.formula(paste("Surv(time, censor) ~", paste(sig_cov, collapse = " + ")))

# Ensure upper scope includes all potential covariates
full_formula <- as.formula(paste("Surv(time, censor) ~", paste(covariates, collapse = " + ")))

# perform forward selection starting from the backward model
fwd_model <- step(
  bkwd,
  scope = list(lower = fwd_formula, upper = full_formula),
  direction = "both",
  k = qchisq(0.10, 1, lower.tail = FALSE)
)
## Start:  AIC=1228.63
## Surv(time, censor) ~ trt + karnof + cd4 + age
##
##               Df    AIC
## + ivdu         1 1228.1
## + black        1 1228.4
## <none>           1228.6
## + cd4_under50  1 1229.8
## + cd450_100    1 1230.0
## + cd4cat       1 1230.7
## + sqrtcd4      1 1230.7
## + hispanic     1 1230.9
## + female       1 1231.2
## + karnof100    1 1231.3
## + cd4_squared  1 1231.3
## + hemophil     1 1231.3
## + priorzdv     1 1231.3
##
## Step:  AIC=1228.06
## Surv(time, censor) ~ trt + karnof + cd4 + age + ivdu
##
##               Df    AIC
## <none>           1228.1
## + black        1 1228.3
## - ivdu         1 1228.6
## + cd4_under50  1 1229.3
## + cd450_100    1 1229.4
## + sqrtcd4      1 1230.1
## + cd4cat       1 1230.1
## + hispanic     1 1230.2
## + female       1 1230.6
## + karnof100    1 1230.6
## + cd4_squared  1 1230.7
## + hemophil     1 1230.7
## + priorzdv     1 1230.8
```

```
# display the final model
summary(fwd_model)
## Call:
## coxph(formula = Surv(time, censor) ~ trt + karnof + cd4 + age +
##     ivdu, data = df)
##
##   n= 1151, number of events= 96
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## trt    -0.671941  0.510716  0.215153 -3.123  0.00179 **
## karnof -0.055725  0.945800  0.011990 -4.647 3.36e-06 ***
## cd4    -0.014509  0.985596  0.002518 -5.762 8.32e-09 ***
## age     0.022148  1.022395  0.011222  1.974  0.04842 *
## ivdu   -0.545165  0.579746  0.322018 -1.693  0.09046 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## trt       0.5107     1.9580    0.3350    0.7786
## karnof    0.9458     1.0573    0.9238    0.9683
## cd4       0.9856     1.0146    0.9807    0.9905
## age       1.0224     0.9781    1.0002    1.0451
## ivdu      0.5797     1.7249    0.3084    1.0898
##
## Concordance= 0.783  (se = 0.023 )
## Likelihood ratio test= 102.4  on 5 df,   p=<2e-16
## Wald test            = 84.84  on 5 df,   p=<2e-16
## Score (logrank) test = 94.46  on 5 df,   p=<2e-16
names(coef(fwd_model))
## [1] "trt"    "karnof" "cd4"    "age"    "ivdu"
```

Listed are the significant covariates from this process. Finally, I will use stepwise regression with significant level 0.10 to prune the main-effects model.

```
sig_cov <- names(coef(fwd_model))
sig_cov <- setdiff(sig_cov, "trt")  # exclude treatment

# create the formula for main effects only
base_formula <- as.formula(
  paste("Surv(time, censor) ~", paste(sig_cov, collapse = " + "), "+ trt")
)

# fit the base model with main effects
base_model <- coxph(base_formula, data = df)
summary(base_model)
## Call:
## coxph(formula = base_formula, data = df)
##
##   n= 1151, number of events= 96
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## karnof -0.055725  0.945800  0.011990 -4.647 3.36e-06 ***
## cd4    -0.014509  0.985596  0.002518 -5.762 8.32e-09 ***
```

```
## age     0.022148  1.022395  0.011222  1.974  0.04842 *
## ivdu   -0.545165  0.579746  0.322018 -1.693  0.09046 .
## trt    -0.671941  0.510716  0.215153 -3.123  0.00179 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## karnof    0.9458     1.0573    0.9238    0.9683
## cd4       0.9856     1.0146    0.9807    0.9905
## age       1.0224     0.9781    1.0002    1.0451
## ivdu      0.5797     1.7249    0.3084    1.0898
## trt       0.5107     1.9580    0.3350    0.7786
##
## Concordance= 0.783  (se = 0.023 )
## Likelihood ratio test= 102.4  on 5 df,   p=<2e-16
## Wald test            = 84.84  on 5 df,   p=<2e-16
## Score (logrank) test = 94.46  on 5 df,   p=<2e-16
```

Here I will add pairwise interaction with treatment by creating interaction terms between treatment and each significant covariate.

```
# interaction terms between treatment and covariates
interaction_terms <- paste("trt *", sig_cov, collapse = " + ")

# Create the formula with main effects + interactions with treatment
itrct_formula <- as.formula(
  paste("Surv(time, censor) ~", paste(sig_cov, collapse = " + "), "+ trt +", interaction_terms)
)

# Fit the model with interactions
interaction_model <- coxph(itrct_formula, data = df)
summary(interaction_model)
## Call:
## coxph(formula = itrct_formula, data = df)
##
##   n= 1151, number of events= 96
##
##                  coef  exp(coef)   se(coef)      z Pr(>|z|)
## karnof     -0.0535520  0.9478566  0.0147334 -3.635 0.000278 ***
## cd4        -0.0144144  0.9856890  0.0031716 -4.545  5.5e-06 ***
## age         0.0172417  1.0173912  0.0143568  1.201 0.229771
## ivdu       -0.4725304  0.6234228  0.3820746 -1.237 0.216180
## trt        -0.6517710  0.5211221  2.5072668 -0.260 0.794900
## karnof:trt -0.0060271  0.9939910  0.0254397 -0.237 0.812721
## cd4:trt    -0.0001825  0.9998175  0.0052235 -0.035 0.972131
## age:trt     0.0130326  1.0131179  0.0230761  0.565 0.572234
## ivdu:trt   -0.2378868  0.7882919  0.7165032 -0.332 0.739881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## karnof        0.9479     1.0550  0.920877    0.9756
## cd4           0.9857     1.0145  0.979581    0.9918
```

```
## age           1.0174     0.9829  0.989162    1.0464
## ivdu          0.6234     1.6040  0.294822    1.3183
## trt           0.5211     1.9189  0.003826   70.9788
## karnof:trt    0.9940     1.0060  0.945645    1.0448
## cd4:trt       0.9998     1.0002  0.989634    1.0101
## age:trt       1.0131     0.9871  0.968317    1.0600
## ivdu:trt      0.7883     1.2686  0.193552    3.2105
##
## Concordance= 0.782  (se = 0.023 )
## Likelihood ratio test= 102.9  on 9 df,   p=<2e-16
## Wald test            = 83.3  on 9 df,   p=4e-14
## Score (logrank) test = 102.5  on 9 df,   p=<2e-16
```

I will use likelihood ratio test to evaluate whether including the interaction tems significantly improves the model.

```
# likelihood ratio test
lr_test <- anova(base_model, interaction_model, test = "LRT")
print(lr_test)
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, censor)
##  Model 1: ~ karnof + cd4 + age + ivdu + trt
##  Model 2: ~ karnof + cd4 + age + ivdu + trt + trt * karnof + trt * cd4 + trt * age + trt * ivdu
##    loglik Chisq Df Pr(>|Chi|)
## 1 -607.26
## 2 -607.00  0.52  4     0.9715
```

The result (p >0.05) indicates that adding the interaction terms does not significantly improve the model's fit ro the data. Therefore, I will cheese `base_model` and perform stepwise regression for the final pruning.

```
# stepwise selection to prune the model
pruned_model <- step(
  base_model,
  direction = "backward",
  k = qchisq(0.10, 1, lower.tail = FALSE)  # Set p-value threshold of 0.10
)
## Start:  AIC=1228.06
## Surv(time, censor) ~ karnof + cd4 + age + ivdu + trt
##
##          Df    AIC
## <none>      1228.1
## - ivdu    1 1228.6
## - age     1 1229.1
## - trt     1 1235.6
## - karnof  1 1246.4
## - cd4     1 1273.4

summary(pruned_model)
## Call:
## coxph(formula = Surv(time, censor) ~ karnof + cd4 + age + ivdu +
##     trt, data = df)
##
##   n= 1151, number of events= 96
```

```
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## karnof -0.055725  0.945800  0.011990 -4.647 3.36e-06 ***
## cd4    -0.014509  0.985596  0.002518 -5.762 8.32e-09 ***
## age     0.022148  1.022395  0.011222  1.974  0.04842 *
## ivdu   -0.545165  0.579746  0.322018 -1.693  0.09046 .
## trt    -0.671941  0.510716  0.215153 -3.123  0.00179 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## karnof    0.9458     1.0573    0.9238    0.9683
## cd4       0.9856     1.0146    0.9807    0.9905
## age       1.0224     0.9781    1.0002    1.0451
## ivdu      0.5797     1.7249    0.3084    1.0898
## trt       0.5107     1.9580    0.3350    0.7786
##
## Concordance= 0.783  (se = 0.023 )
## Likelihood ratio test= 102.4  on 5 df,   p=<2e-16
## Wald test            = 84.84  on 5 df,   p=<2e-16
## Score (logrank) test = 94.46  on 5 df,   p=<2e-16
```

All the covariates meet p-value <0.10.

```
# table of -2logL and AIC for the various models I considered in the process of model selection
data.frame(
  Model = c("Base Model", "Interaction Model", "Pruned Model"),
  Minus2LogL = c(
    -2 * logLik(base_model),
    -2 * logLik(interaction_model),
    -2 * logLik(pruned_model)
  ),
  AIC = c(
    AIC(base_model),
    AIC(interaction_model),
    AIC(pruned_model)
  )
) |> kable()
```

Model	Minus2LogL	AIC
Base Model	1214.528	1224.528
Interaction Model	1214.008	1232.008
Pruned Model	1214.528	1224.528

We can confirm that the pruned model has the lowest AIC and should be used as the final model.

```
# summary table of the final model
final_summary <- summary(pruned_model)

# create a data frame for parameter estimates
data.frame(
```

```
  Coefficient = final_summary$coefficients[, "coef"],
  SE = final_summary$coefficients[, "se(coef)"],
  "P-value" = final_summary$coefficients[, "Pr(>|z|)"],
  HR = exp(final_summary$coefficients[, "coef"])
) |>
  kable()
```

	Coefficient	SE	P.value	HR
karnof	-0.0557246	0.0119904	0.0000034	0.9457996
cd4	-0.0145090	0.0025181	0.0000000	0.9855958
age	0.0221483	0.0112220	0.0484214	1.0223954
ivdu	-0.5451651	0.3220175	0.0904618	0.5797461
trt	-0.6719408	0.2151534	0.0017897	0.5107164

Treatment with IDV is significantly associated with a reduced risk of AIDS progression or death. Patients receiving IDV experience approximately a 49% reduction in the hazard compared to those not receiving IDV, holding other variables constant. A higher Karnofsky score (better functional status) is significantly associated with a reduced risk of AIDS progression or death. For every 1-point increase in the Karnofsky score, the hazard decreases by approximately 5.4%, holding other variables constant. A higher baseline CD4 count is significantly associated with a reduced risk of AIDS progression or death. For every 1-cell increase in baseline CD4 count, the hazard decreases by approximately 1.4%, holding other variables constant. Age is also a significant predictor. Older age is associated with a slightly increased risk of AIDS progression or death, with a hazard increase of approximately 2.2% per year, holding other variables constant. Prior IV drug use appears to be associated with a reduced hazard, though the association is not statistically significant at the 0.05 level.
The treatment effect alone (unadjusted) shows a significant reduction in the risk of AIDS progression or death. After adjusting for covariates such as Karnofsky score, CD4 count, age, and IV drug use, the treatment effect remains significant. This indicates that the effect of treatment is not heavily confounded by these variables.

2. Evaluating Linearity of CD4 Effects

a. CD4 as continuous, on original scale and square root transformation

```
# original scale
cd4_cont <- coxph(Surv(time, censor) ~ cd4 + age + ivdu + karnof + trt, data = df)

# square root transformation
df$sqrt_cd4 <- sqrt(df$cd4)
cd4_sqrt <- coxph(Surv(time, censor) ~ sqrt_cd4 + age + ivdu + karnof + trt, data = df)
```

b. CD4 as ordinal (with levels 1=<50, 2=50-100, and 3=> 100.)

```
df$cd4_ordinal <- cut(df$cd4, breaks = c(-Inf, 50, 100, Inf), labels = c(1, 2, 3), right = FALSE)
cd4_ordinal <- coxph(Surv(time, censor) ~ cd4_ordinal + age + ivdu + karnof + trt, data = df)
```

c. CD4 as categorical (create indicators for <50 and 50-100)

```r
df$cd4_under50 <- ifelse(df$cd4 < 50, 1, 0)
df$cd4_50_100 <- ifelse(df$cd4 >= 50 & df$cd4 <= 100, 1, 0)
cd4_cat <- coxph(Surv(time, censor) ~ cd4_under50 + cd4_50_100 + age + ivdu + karnof + trt, data = df)
```

d. CD4 as linear and quadratic terms (cd4 and cd4^2)

```r
df$cd4_squared <- df$cd4^2
cd4_quad <- coxph(Surv(time, censor) ~ cd4 + cd4_squared + age + ivdu + karnof + trt, data = df)
```

```r
# summary table for comparison
data.frame(
  Model = c("Continuous (original)", "Square Root", "Ordinal", "Categorical", "Linear + Quadratic"),
  Minus2LogL = c(
    -2 * logLik(cd4_cont),
    -2 * logLik(cd4_sqrt),
    -2 * logLik(cd4_ordinal),
    -2 * logLik(cd4_cat),
    -2 * logLik(cd4_quad)
  ),
  AIC = c(
    AIC(cd4_cont),
    AIC(cd4_sqrt),
    AIC(cd4_ordinal),
    AIC(cd4_cat),
    AIC(cd4_quad)
  )
) |> kable()
```

Model	Minus2LogL	AIC
Continuous (original)	1214.528	1224.528
Square Root	1220.741	1230.741
Ordinal	1223.194	1235.194
Categorical	1221.465	1233.465
Linear + Quadratic	1214.472	1226.472

The Continuous (original) model for CD4 is preferred as it provides the best prediction based on AIC and simplicity, while maintaining comparable fit to the data. The Linear + Quadratic model, while slightly more flexible, does not justify its added complexity given the small difference in fit.

3. Predicted Time to AIDS or Death Under PH and KM

4. Assessing the Proportional Hazards Assumption

a.
b.
c.

5. Evaluating Fit of the Final Model