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

```r
# define CD4 category as in part 2b
df$cd4_cat <- cut(df$cd4, breaks = c(-Inf, 50, 100, Inf), labels = c("1", "2", "3"), right = FALSE)

# define Karnofsky score category
df$karnof100 <- ifelse(df$karnof == 100, 1, 0)
```

I will fit the cox model with treatment, CD4 category, and Karnofsky score (100 vs. lower).

```r
cox_model <- coxph(Surv(time, censor) ~ cd4_cat + trt + karnof100, data = df)

summary(cox_model)
## Call:
## coxph(formula = Surv(time, censor) ~ cd4_cat + trt + karnof100,
##     data = df)
##
##   n= 1151, number of events= 96
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## cd4_cat2  -0.6013    0.5481   0.2601 -2.312  0.02079 *
## cd4_cat3  -1.9514    0.1421   0.3551 -5.496 3.88e-08 ***
## trt       -0.6733    0.5100   0.2152 -3.129  0.00175 **
## karnof100 -0.8123    0.4438   0.2683 -3.027  0.00247 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## cd4_cat2     0.5481      1.824   0.32922    0.9126
## cd4_cat3     0.1421      7.039   0.07084    0.2849
## trt          0.5100      1.961   0.33454    0.7776
## karnof100    0.4438      2.253   0.26233    0.7509
##
## Concordance= 0.737  (se = 0.023 )
## Likelihood ratio test= 71.83  on 4 df,   p=9e-15
## Wald test            = 52.95  on 4 df,   p=9e-11
## Score (logrank) test = 64.81  on 4 df,   p=3e-13
```

Now, let's calculate the predicted probability of remaining alive and AIDS-free for each combination of these covariates at 6 months and 1 year.

```r
# create a new data frame for prediction
pred_dat <- expand.grid(
  time = c(6, 12),
  censor = c(0, 1),
  cd4_cat = c("1", "2", "3"),
  trt = c(0, 1),
  karnof100 = c(0, 1)
)

# predict survival probabilities
preds <- predict(cox_model, newdata = pred_dat, type = "expected", se.fit = TRUE)

# convert predicted survival probabilities
pred_dat$prob <- exp(-preds$fit) # survival probability: exp(-expected risk)
print(pred_dat)
##    time censor cd4_cat trt karnof100      prob
## 1     6      0       1   0         0 0.8058679
## 2    12      0       1   0         0 0.7425467
## 3     6      1       1   0         0 0.8058679
## 4    12      1       1   0         0 0.7425467
## 5     6      0       2   0         0 0.8884264
## 6    12      0       2   0         0 0.8494567
```

```
## 7     6      1       2   0         0 0.8884264
## 8    12      1       2   0         0 0.8494567
## 9     6      0       3   0         0 0.9698012
## 10   12      0       3   0         0 0.9585913
## 11    6      1       3   0         0 0.9698012
## 12   12      1       3   0         0 0.9585913
## 13    6      0       1   1         0 0.8957601
## 14   12      0       1   1         0 0.8591425
## 15    6      1       1   1         0 0.8957601
## 16   12      1       1   1         0 0.8591425
## 17    6      0       2   1         0 0.9414459
## 18   12      0       2   1         0 0.9201527
## 19    6      1       2   1         0 0.9414459
## 20   12      1       2   1         0 0.9201527
## 21    6      0       3   1         0 0.9844820
## 22   12      0       3   1         0 0.9786615
## 23    6      1       3   1         0 0.9844820
## 24   12      1       3   1         0 0.9786615
## 25    6      0       1   0         1 0.9086483
## 26   12      0       1   0         1 0.8762370
## 27    6      1       1   0         1 0.9086483
## 28   12      1       1   0         1 0.8762370
## 29    6      0       2   0         1 0.9488465
## 30   12      0       2   0         1 0.9301432
## 31    6      1       2   0         1 0.9488465
## 32   12      1       2   0         1 0.9301432
## 33    6      0       3   0         1 0.9864821
## 34   12      0       3   0         1 0.9814047
## 35    6      1       3   0         1 0.9864821
## 36   12      1       3   0         1 0.9814047
## 37    6      0       1   1         1 0.9523150
## 38   12      0       1   1         1 0.9348357
## 39    6      1       1   1         1 0.9523150
## 40   12      1       1   1         1 0.9348357
## 41    6      0       2   1         1 0.9735746
## 42   12      0       2   1         1 0.9637390
## 43    6      1       2   1         1 0.9735746
## 44   12      1       2   1         1 0.9637390
## 45    6      0       3   1         1 0.9930825
## 46   12      0       3   1         1 0.9904722
## 47    6      1       3   1         1 0.9930825
## 48   12      1       3   1         1 0.9904722
```

```
# calculate survival estimates under KM
# fit KM survival curves stratified by CD4 category, treatment, and Karnofsky score
pred_dat <- expand.grid(
  time = c(6, 12),
  censor = c(0, 1),
  cd4_cat = c("1", "2", "3"),
  trt = c(0, 1),
  karnof100 = c(0, 1)
)
```

```
km_fit <- survfit(Surv(time, censor) ~ cd4_cat + trt + karnof100, data = df)
summary(km_fit, 6)
## Call: survfit(formula = Surv(time, censor) ~ cd4_cat + trt + karnof100,
##     data = df)
##
##                 cd4_cat=1, trt=0, karnof100=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000     108.0000      32.0000       0.7946       0.0325       0.7333
## upper 95% CI
##       0.8609
##
##                 cd4_cat=1, trt=0, karnof100=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000      58.0000       3.0000       0.9541       0.0259       0.9047
## upper 95% CI
##       1.0000
##
##                 cd4_cat=1, trt=1, karnof100=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000     126.0000      17.0000       0.8922       0.0247       0.8450
## upper 95% CI
##       0.9420
##
##                 cd4_cat=1, trt=1, karnof100=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000      48.0000       3.0000       0.9455       0.0306       0.8873
## upper 95% CI
##       1.0000
##
##                 cd4_cat=2, trt=0, karnof100=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000      49.0000       6.0000       0.9152       0.0334       0.8521
## upper 95% CI
##       0.9831
##
##                 cd4_cat=2, trt=0, karnof100=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000      34.0000       4.0000       0.9139       0.0415       0.8361
## upper 95% CI
##       0.9990
##
##                 cd4_cat=2, trt=1, karnof100=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000      53.0000       4.0000       0.9389       0.0297       0.8825
## upper 95% CI
##       0.9989
##
##                 cd4_cat=2, trt=1, karnof100=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000      32.0000       2.0000       0.9545       0.0315       0.8948
## upper 95% CI
##       1.0000
##
```

```
##                 cd4_cat=3, trt=0, karnof100=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000      94.0000       3.0000       0.9748       0.0144       0.9471
## upper 95% CI
##       1.0000
##
##                 cd4_cat=3, trt=0, karnof100=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000      60.0000       2.0000       0.9696       0.0212       0.9290
## upper 95% CI
##       1.0000
##
##                 cd4_cat=3, trt=1, karnof100=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     6.00e+00     1.09e+02     1.00e+00     9.93e-01     7.33e-03     9.78e-01
## upper 95% CI
##     1.00e+00
##
##                 cd4_cat=3, trt=1, karnof100=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000      64.0000       1.0000       0.9891       0.0108       0.9682
## upper 95% CI
##       1.0000
summary(km_fit, 12)
## Call: survfit(formula = Surv(time, censor) ~ cd4_cat + trt + karnof100,
##     data = df)
##
##                 cd4_cat=1, trt=0, karnof100=0
##      time n.risk n.event survival
##
##                 cd4_cat=1, trt=0, karnof100=1
##      time n.risk n.event survival
##
##                 cd4_cat=1, trt=1, karnof100=0
##      time n.risk n.event survival
##
##                 cd4_cat=1, trt=1, karnof100=1
##      time n.risk n.event survival
##
##                 cd4_cat=2, trt=0, karnof100=0
##      time n.risk n.event survival
##
##                 cd4_cat=2, trt=0, karnof100=1
##      time n.risk n.event survival
##
##                 cd4_cat=2, trt=1, karnof100=0
##      time n.risk n.event survival
##
##                 cd4_cat=2, trt=1, karnof100=1
##      time n.risk n.event survival
##
##                 cd4_cat=3, trt=0, karnof100=0
##      time n.risk n.event survival
```

```
##
##                 cd4_cat=3, trt=0, karnof100=1
##      time n.risk n.event survival
##
##                 cd4_cat=3, trt=1, karnof100=0
##      time n.risk n.event survival
##
##                 cd4_cat=3, trt=1, karnof100=1
##      time n.risk n.event survival
```

The summary table is shown below.

			Estimates Under PH		Kaplan-Meier Estimates	
CD4 category	Karnofsky	Treatment	$\hat{S}(6)$	$\hat{S}(12)$	$\hat{S}(6)$	$\hat{S}(12)$
<50	70-90	no IDV	0.81	0.74	0.79	NA
	70-90	IDV	0.90	0.86	0.89	NA
	100	no IDV	0.91	0.88	0.95	NA
	100	IDV	0.95	0.93	0.95	NA
50-100	70-90	no IDV	0.89	0.85	0.92	NA
	70-90	IDV	0.94	0.92	0.94	NA
	100	no IDV	0.91	0.88	0.91	NA
	100	IDV	0.95	0.93	0.95	NA
>100	70-90	no IDV	0.97	0.96	0.97	NA
	70-90	IDV	0.98	0.98	0.99	NA
	100	no IDV	0.99	0.98	0.97	NA
	100	IDV	0.99	0.99	0.99	NA

The baseline is CD4 category <50, Karnofsky 70-90, and treatment without IDV. The table shows that subgroup with CD4 <50, Karnofsky 70-90, treatment without IDV has the highest risk of AIDS or death at 12 months.

To calculate the 6-month survival probability for a specific group using the predicted survival of a baseline group in the Cox PH model, I will use the hazard ratios from the model.

The survival probability for a specific group is given by: $S(6) = S_0(6)^{\exp(X\beta)}$

Where:

- $S(6)$: Survival probability at 6 months
- $S_0(6)$: Baseline survival probability at 6 months
- X: Covariate vector
- β: Coefficient vector from the Cox model

```
# baseline survival probability at 6 month
summary(survfit(cox_model), times = 6)$surv
## [1] 0.8058679
```

Since subgroup with CD4 <50, Karnofsky 70-90, treatment without IDV is the baseline group, $\exp(X\beta) = e^0 = 1$. Therefore, $S(6) = 0.8058679^1 \approx 0.81$. This is what we have in the table.

The KM method shows slightly lower survival probabilities at 6 months for subgroups with lower Karnofsky scores or no IDV treatment compared to PH estimates. This suggests that the proportional hazards

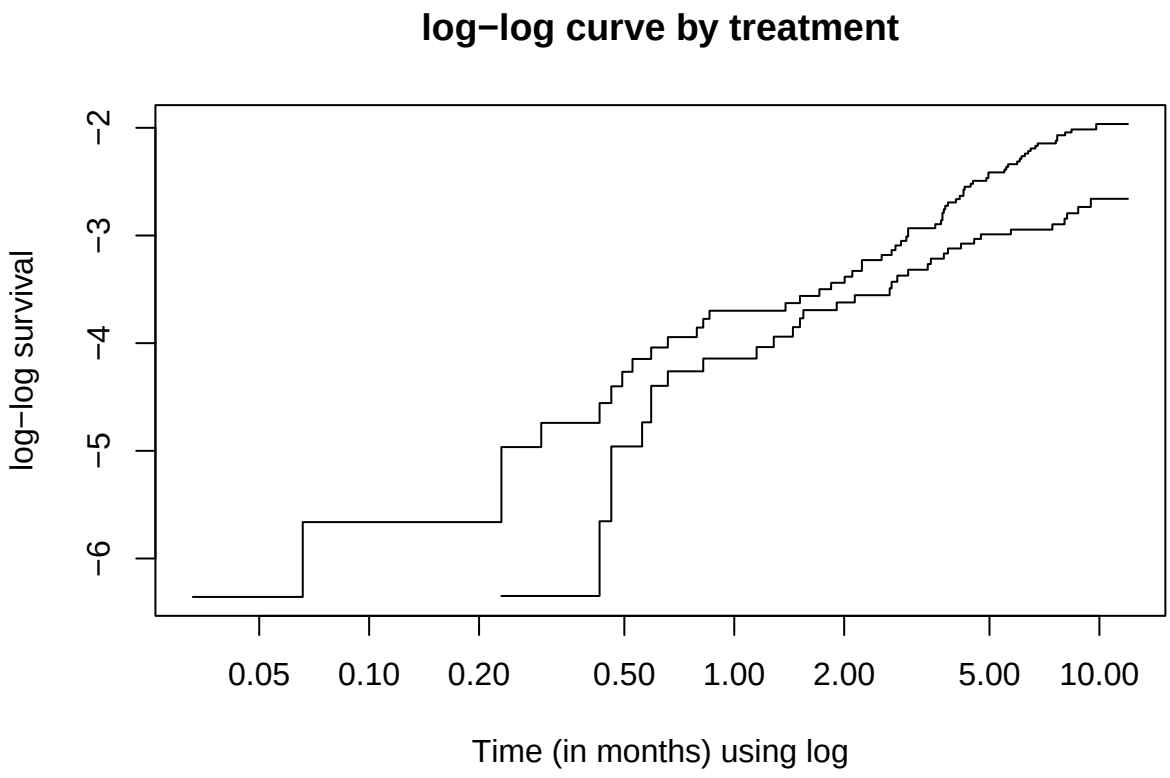
assumption in the Cox model may overestimate survival for some groups. At 12 months, KM estimates are unavailable for many subgroups due to insufficient follow-up data, while PH estimates can extrapolate survival probabilities based on model assumptions. Both methods consistently show better survival with higher CD4 count, IDV treatment and higher Karnofsky scores.

4. Assessing the Proportional Hazards Assumption

The effect of treatment on AIDS-free survival

a. Creating a plot of log[−log(Sˆ(t)] versus log(t) for each treatment arm

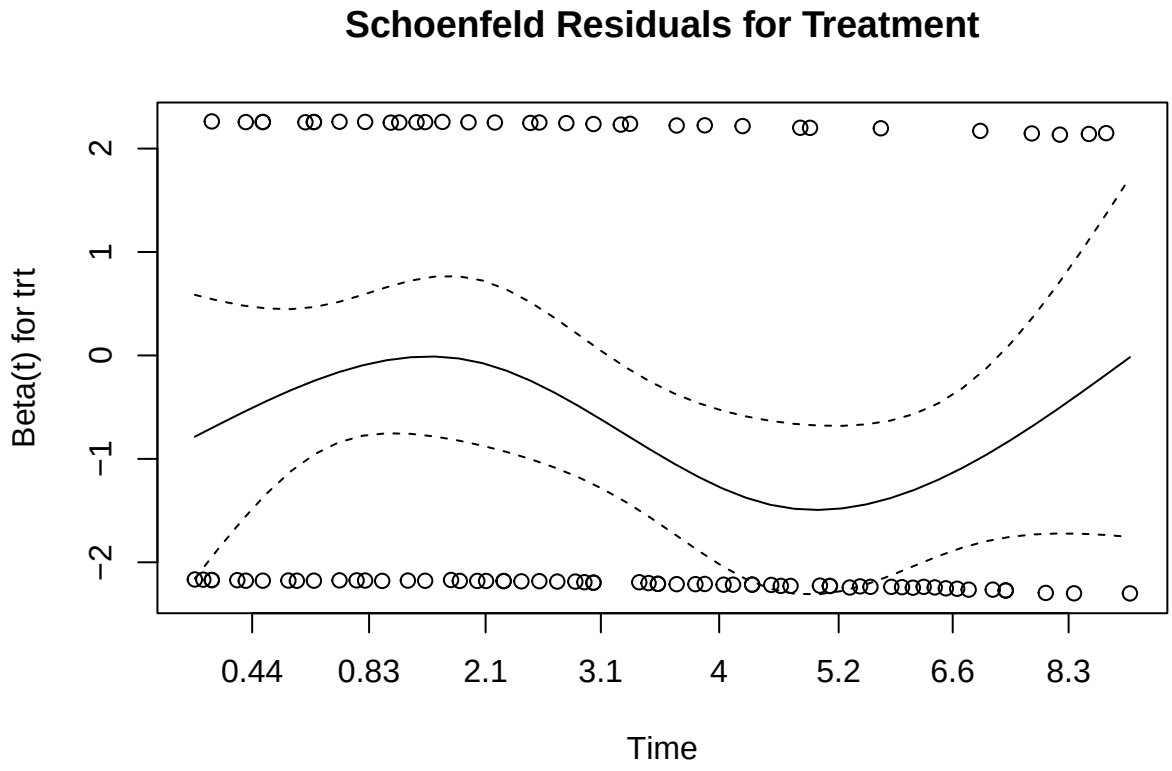
```
# fit Kaplan-Meier survival curves stratified by treatment
km_fit <- survfit(Surv(time, censor) ~ trt, data = df)

# plot log[-log(S(t))] vs log(t)
plot(km_fit, fun = "cloglog", xlab = "Time (in months) using log",
     ylab = "log-log survival", main = "log-log curve by treatment")
```



The curves appear to be non-parallel, suggesting a potential violation of the PH assumption.

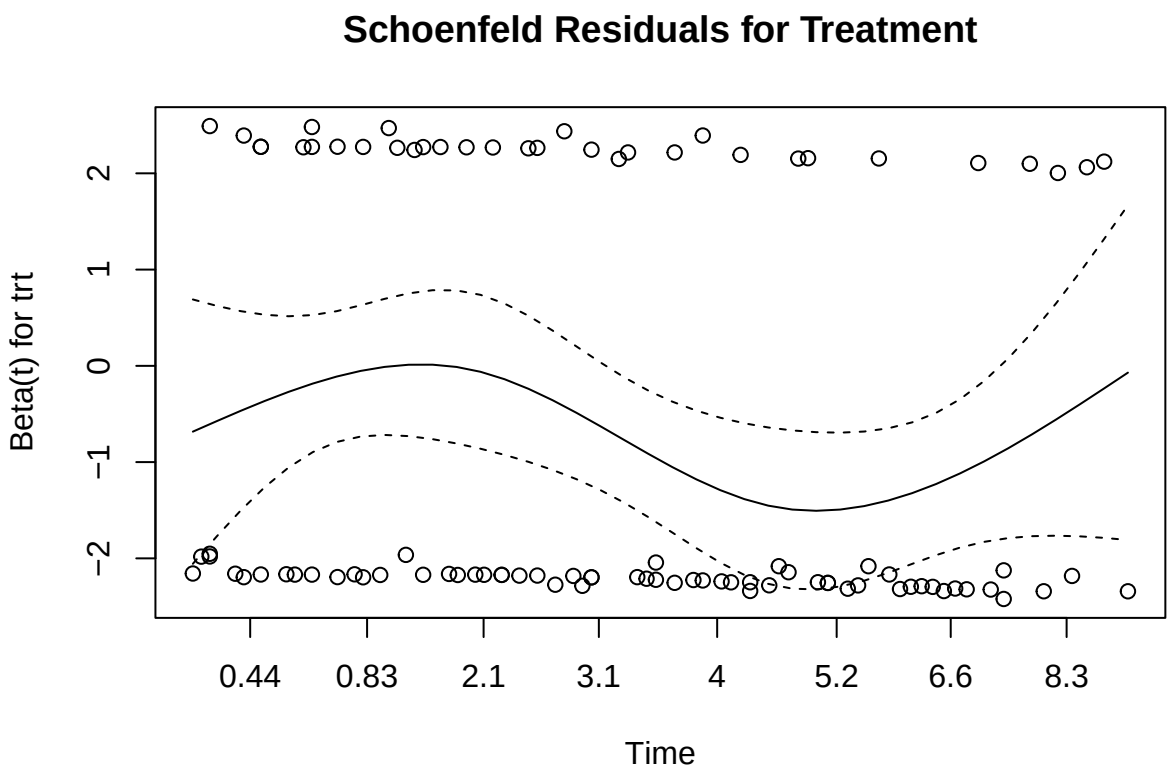
b. Evaluating weighted Schoenfeld residuals over time

```r
# test the proportional hazards assumption
cox_model <- coxph(Surv(time, censor) ~ trt , data = df)
cox_zph <- cox.zph(cox_model)

# summary of Schoenfeld residuals
print(cox_zph)
##        chisq df   p
## trt     1.66  1 0.2
## GLOBAL  1.66  1 0.2

# plot Schoenfeld residuals for treatment
plot(cox_zph, var = "trt", main = "Schoenfeld Residuals for Treatment")
```




```r
# test the proportional hazards assumption (with adjustment)
cox_model <- coxph(Surv(time, censor) ~ trt + cd4_cat + karnof100, data = df)
cox_zph <- cox.zph(cox_model)

# summary of Schoenfeld residuals
print(cox_zph)
##            chisq df    p
## trt         1.95  1 0.16
## cd4_cat     1.84  2 0.40
## karnof100   1.30  1 0.25
## GLOBAL      5.45  4 0.24
```

```r
# plot Schoenfeld residuals for treatment
plot(cox_zph, var = "trt", main = "Schoenfeld Residuals for Treatment")
```



The both plots there appears to be a trend in residuals over time, which indicates that the effect of treatment may vary with time. Hence, PH assumption may be violated.

c. Including an interaction with time and treatment

```r
# add interaction term between treatment and time
cox_int <- coxph(Surv(time, censor) ~ trt * time + cd4_cat + karnof100, data = df)

# summary of the model with interaction
summary(cox_int)
## Call:
## coxph(formula = Surv(time, censor) ~ trt * time + cd4_cat + karnof100,
##     data = df)
##
##   n= 1151, number of events= 96
##
##                 coef  exp(coef)   se(coef)      z Pr(>|z|)
## trt         6.180e-02  1.064e+00  2.914e-01  0.212 0.832044
## time       -9.287e+01  4.646e-41  1.471e+01 -6.314 2.72e-10 ***
## cd4_cat2   -6.726e-01  5.104e-01  3.493e-01 -1.926 0.054149 .
## cd4_cat3   -1.450e+00  2.346e-01  4.071e-01 -3.562 0.000369 ***
## karnof100 -7.453e-01  4.746e-01  3.872e-01 -1.925 0.054284 .
```

```
## trt:time  -7.401e-02  9.287e-01  7.374e-02 -1.004 0.315566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## trt       1.064e+00  9.401e-01 6.009e-01 1.883e+00
## time      4.646e-41  2.152e+40 1.402e-53 1.540e-28
## cd4_cat2  5.104e-01  1.959e+00 2.574e-01 1.012e+00
## cd4_cat3  2.346e-01  4.262e+00 1.057e-01 5.210e-01
## karnof100 4.746e-01  2.107e+00 2.222e-01 1.014e+00
## trt:time  9.287e-01  1.077e+00 8.037e-01 1.073e+00
##
## Concordance= 1  (se = 0 )
## Likelihood ratio test= 1184  on 6 df,   p=<2e-16
## Wald test            = 61.01  on 6 df,   p=3e-11
## Score (logrank) test = 585.7  on 6 df,   p=<2e-16
```

Including an interaction term between treatment and time in the Cox model provided a hazard ratio for the interaction ($HR = 0.91$), indicating a slight decrease in the treatment effect over time, but this result was not statistically significant. It suggests weak evidence of time-dependency in the treatment effect.

```
km_fit <- survfit(Surv(time, censor) ~ trt, data = df)
summary(km_fit, 3)
## Call: survfit(formula = Surv(time, censor) ~ trt, data = df)
##
##                 trt=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     3.00e+00     5.04e+02     2.90e+01     9.48e-01     9.39e-03     9.30e-01
## upper 95% CI
##     9.67e-01
##
##                 trt=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     3.00e+00     5.17e+02     2.00e+01     9.64e-01     7.82e-03     9.49e-01
## upper 95% CI
##     9.80e-01
summary(km_fit, 6)
## Call: survfit(formula = Surv(time, censor) ~ trt, data = df)
##
##                 trt=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       6.0000     403.0000      50.0000       0.9057       0.0127       0.8811
## upper 95% CI
##       0.9311
##
##                 trt=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     6.00e+00     4.32e+02     2.80e+01     9.49e-01     9.45e-03     9.30e-01
## upper 95% CI
##     9.67e-01
summary(km_fit, 9)
## Call: survfit(formula = Surv(time, censor) ~ trt, data = df)
##
```

```
##                 trt=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##       9.0000     211.0000      62.0000       0.8752       0.0151       0.8461
## upper 95% CI
##       0.9053
##
##                 trt=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##        9.000      242.000       32.000        0.937        0.011        0.916
## upper 95% CI
##        0.959

# HR
-log(0.964) / -log(0.948) # at 3 month
## [1] 0.6865815
-log(0.949) / -log(0.9057) # at 6 month
## [1] 0.5285006
-log(0.937) / -log(0.8752) # at 9 month
## [1] 0.4881516
```

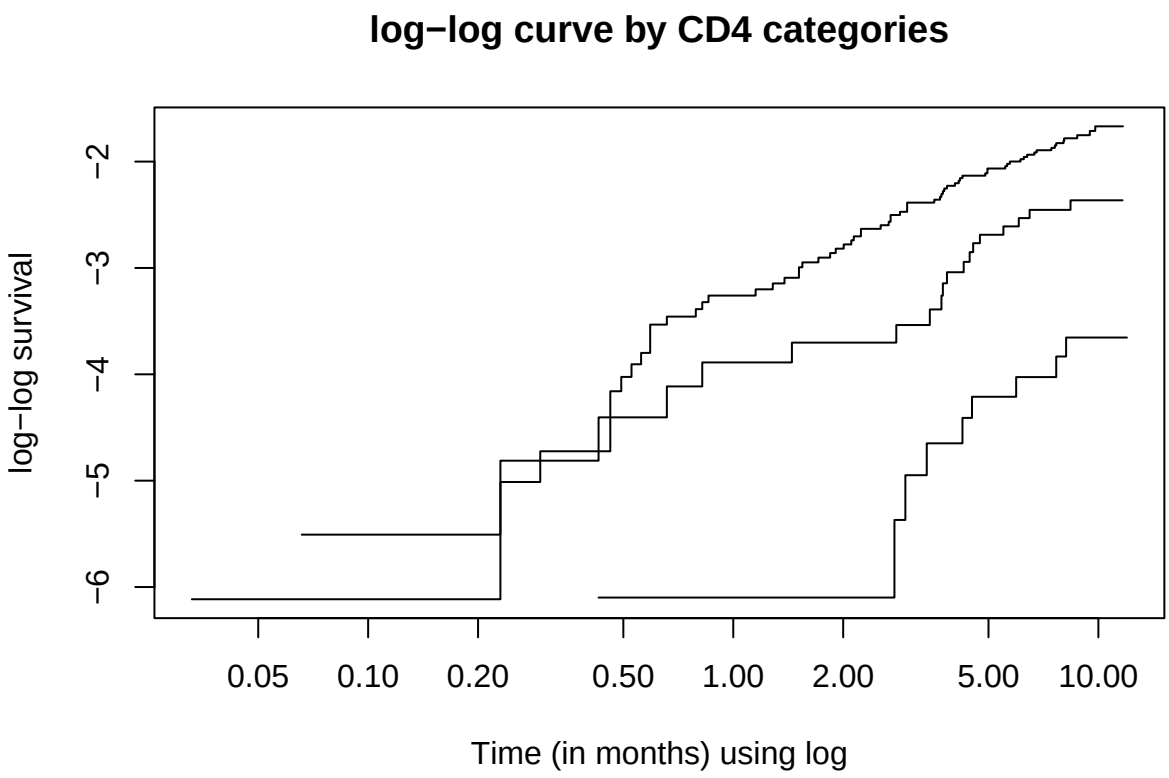
The predicted HR for 3, 6, and 9 months are 0.69, 0.53, and 0.49, accordingly.

Overall, (a)-(c) imply that the PH assumption is violated.

The effect of CD4 count on AIDS-free survival

a. Creating a plot of log[−log(Sˆ(t)] versus log(t) for each treatment arm

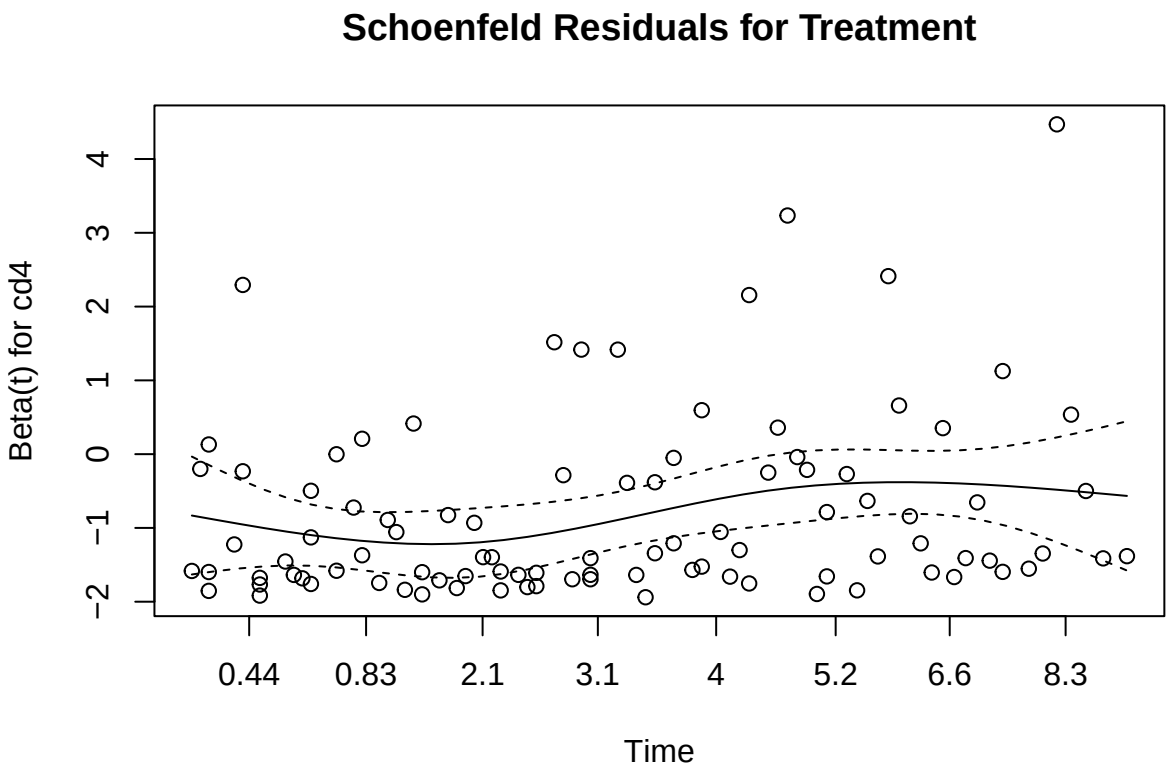
```
# fit Kaplan-Meier survival curves stratified by treatment
km_fit <- survfit(Surv(time, censor) ~ cd4_cat, data = df)

# plot log[-log(S(t))] vs log(t)
plot(km_fit, fun = "cloglog", xlab = "Time (in months) using log",
     ylab = "log-log survival", main = "log-log curve by CD4 categories")
```



The curves appear to be non-parallel, suggesting a potential violation of the PH assumption.

b. Evaluating weighted Schoenfeld residuals over time

```
df_cd4 = df |>
  mutate(
    cd4 = cd4/50 # divide cd4 by 50
  )

# test the proportional hazards assumption
cox_model <- coxph(Surv(time, censor) ~ cd4, data = df_cd4)
cox_zph <- cox.zph(cox_model)

# summary of Schoenfeld residuals
print(cox_zph)
##        chisq df     p
## cd4     4.28  1 0.039
## GLOBAL  4.28  1 0.039

# plot Schoenfeld residuals for treatment
plot(cox_zph, var = "cd4", main = "Schoenfeld Residuals for Treatment")
```



There appears to be a trend in residuals over time, which indicates that the effect of treatment may vary with time. PH assumption may be violated.

c. Including an interaction with time and treatment

```
# add interaction term between treatment and time
cox_int <- coxph(Surv(time, censor) ~ cd4 * trt, data = df)

# summary of the model with interaction
summary(cox_int)
## Call:
## coxph(formula = Surv(time, censor) ~ cd4 * trt, data = df)
##
##   n= 1151, number of events= 96
##
##              coef exp(coef)  se(coef)      z Pr(>|z|)
## cd4     -0.016262  0.983870  0.003166 -5.136 2.81e-07 ***
## trt     -0.660845  0.516415  0.290386 -2.276   0.0229 *
## cd4:trt  0.000366  1.000366  0.005207  0.070   0.9440
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## cd4        0.9839     1.0164    0.9778    0.9900
## trt        0.5164     1.9364    0.2923    0.9124
## cd4:trt    1.0004     0.9996    0.9902    1.0106
```

```
##
## Concordance= 0.744  (se = 0.022 )
## Likelihood ratio test= 73.24  on 3 df,   p=9e-16
## Wald test            = 50.85  on 3 df,   p=5e-11
## Score (logrank) test = 62.96  on 3 df,   p=1e-13
```

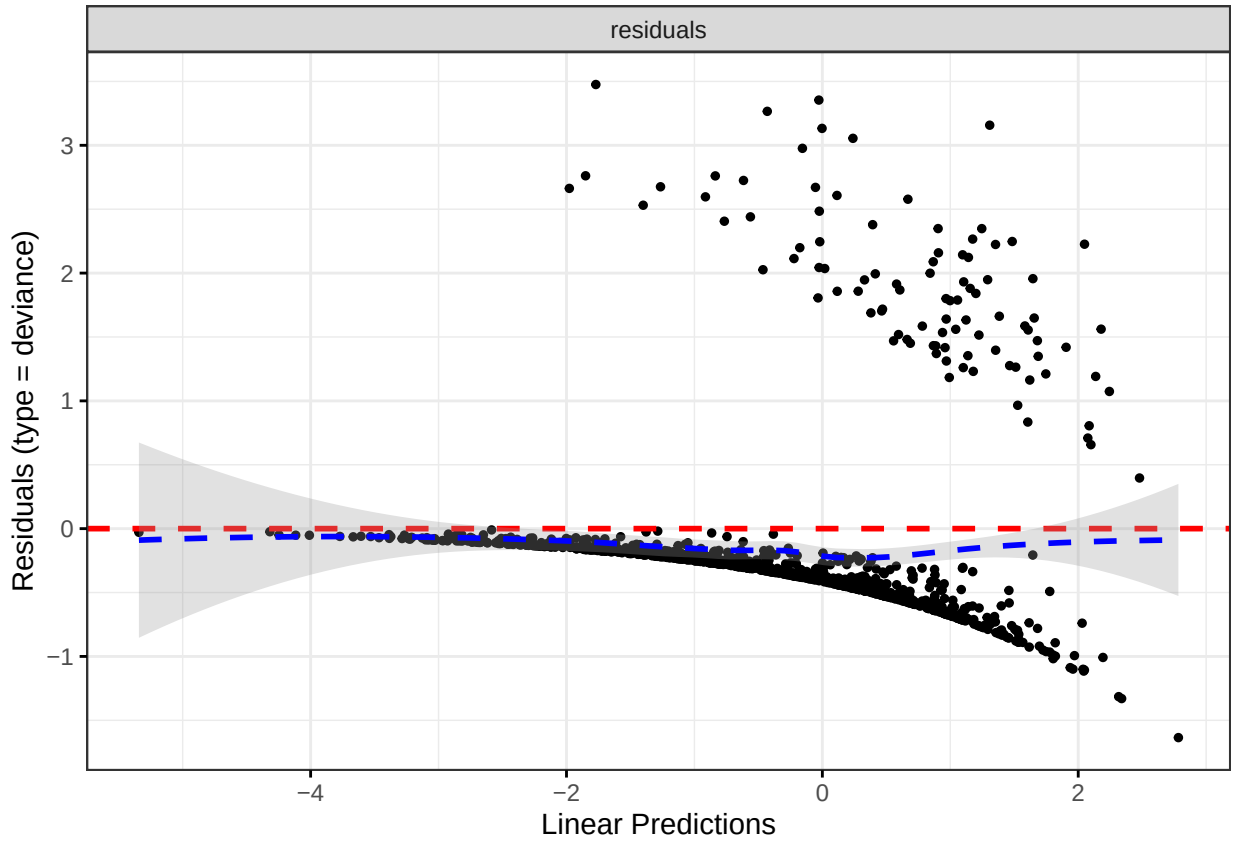
Interaction term is not statistically significant, suggesting a weak evidence of treatment-dependency in the CD4 count.

(a)-(c) imply that the PH assumption is violated.

5. Evaluating Fit of the Final Model

Deviance residuals

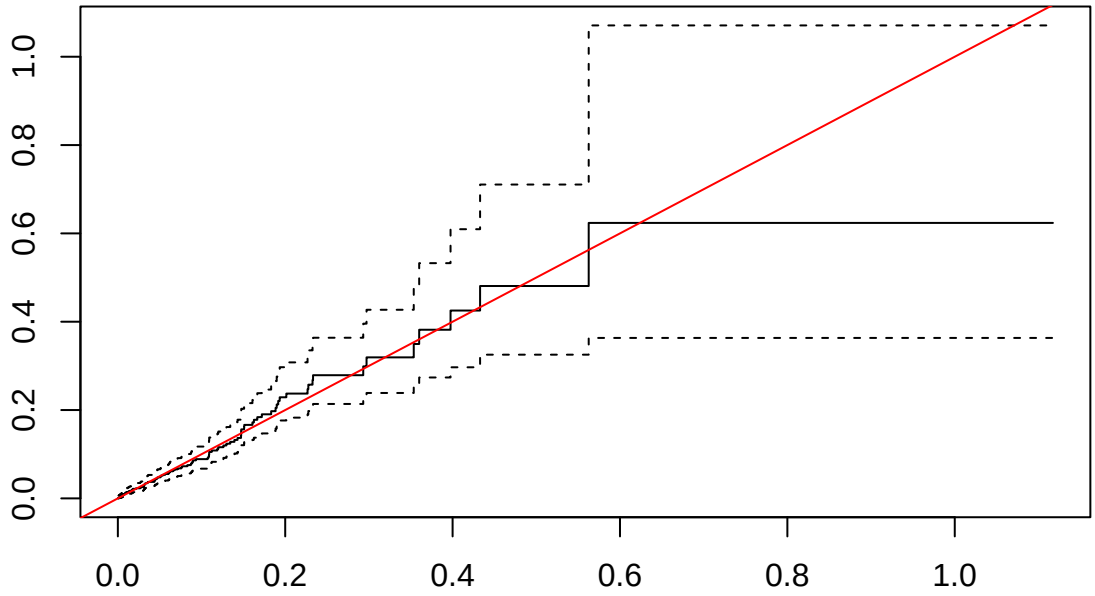
```
cox_model <- coxph(Surv(time, censor) ~ trt + karnof + cd4 + age + ivdu, data = df)

ggcoxdiagnostics(
  cox_model,
  type = c("deviance")
  )
```



Deviance residuals plot shows some asymmetry and a few outliers on the higher end of the residual scale. This indicates that some individuals may not align perfectly with the assumptions of the model, but the majority of the residuals cluster around zero, suggesting an overall reasonable fit.

Cox-Snell generalized residuals

```
fitg <- flexsurvreg(formula = Surv(time, censor) ~ trt + karnof + cd4 + age + ivdu,
                    data = df, dist = "gengamma")
cs <- coxsnell_flexsurvreg(fitg)

surv <- survfit(Surv(cs$est, df$censor) ~ 1)
plot(surv, fun="cumhaz")
abline(0, 1, col="red")
```

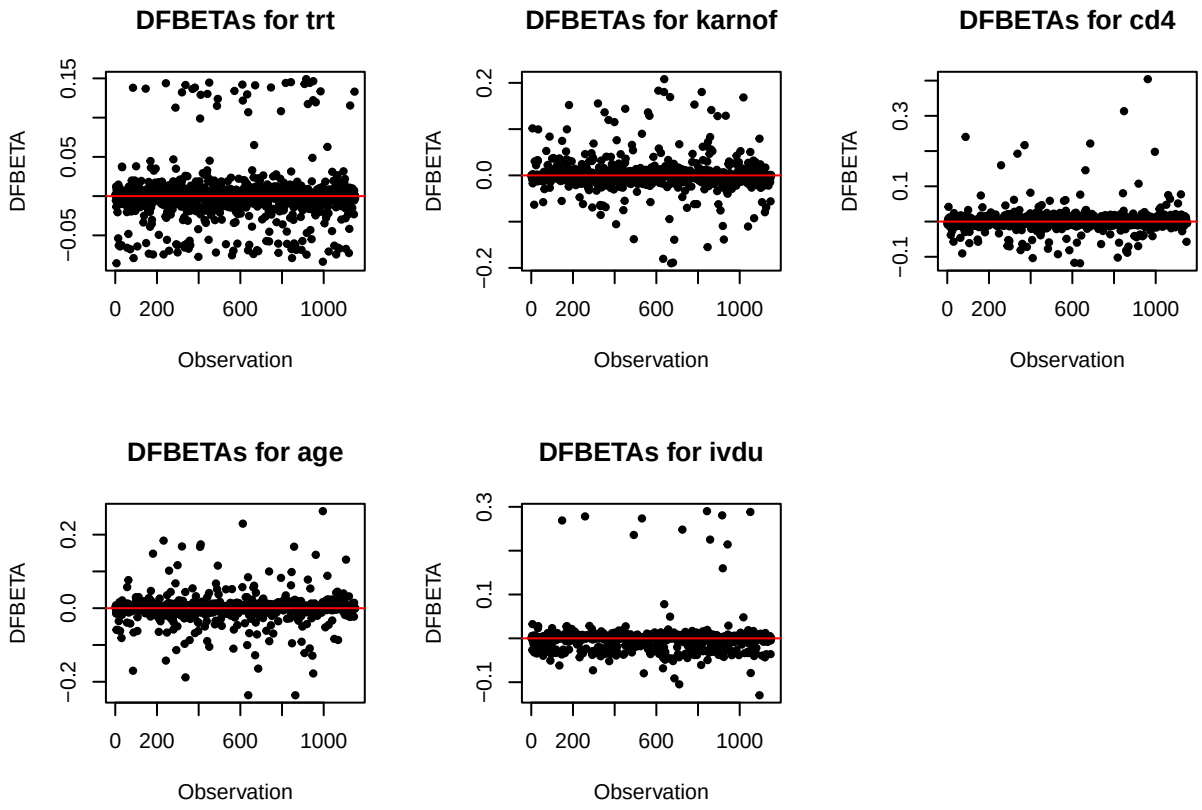


The Cox-Snell residuals follow the expected cumulative hazard reasonably well, although there are some deviations at higher survival probabilities.

Influence diagnostics

```
# extract influence measures
influence_measures <- cox.zph(cox_model, transform = "identity")

# calculate DFBETAs
dfbetas <- residuals(cox_model, type = "dfbetas")
colnames(dfbetas) <- c("trt", "karnof", "cd4", "age", "ivdu")

# plot DFBETAs for each coefficient
```

```r
par(mfrow = c(2, 3))
for (i in 1:ncol(dfbetas)) {
  plot(dfbetas[, i], main = paste("DFBETAs for", colnames(dfbetas)[i]),
       xlab = "Observation", ylab = "DFBETA", pch = 20)
  abline(h = 0, col = "red")
}
par(mfrow = c(1, 1))
```



The DFBETAs indicate that certain observations (outliers) have substantial influence on specific coefficients such as cd4 and ivdu. These patients could potentially distort the model results.

In conclusion, the model provides a reasonable fit to the data, but the presence of influential observations suggests that the final model may be sensitive to these outliers.