CoxPHModel

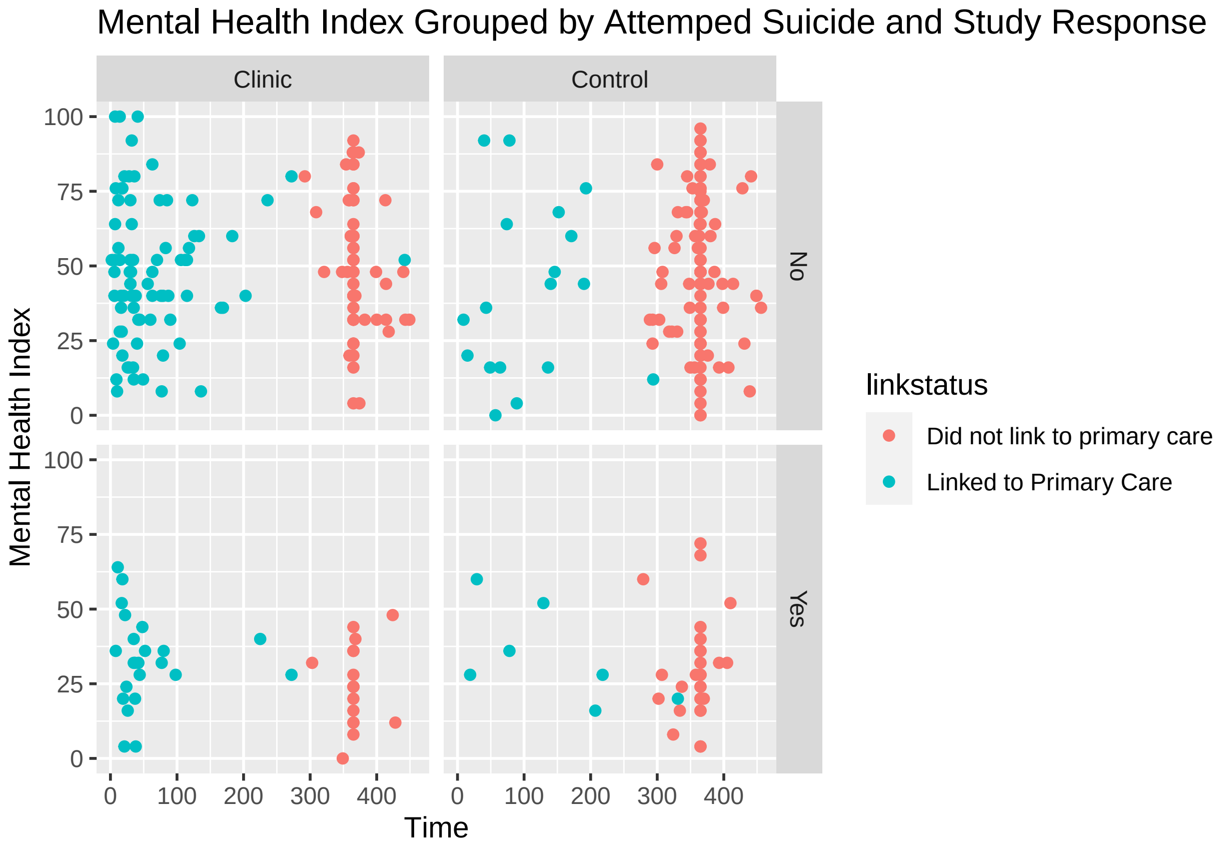
```r
library(mosaic)
library(readr)
library(tidyverse)
library(broom)
library(survival)
library(survminer)
library(praise)
```

```r
df <- read_csv("HELPdata.csv", na="*")
```

```r
df <- df %>%
  mutate(yrs_education = as.numeric(a9),
         gender=a1,
         alcq_30 = as.numeric(alcq_30),
         marriage = as.factor(a10),
         employment = as.factor(a13),
         income = as.factor(case_when(a18 == 1 ~ "<5000",
                                      a18 == 2 ~ "5000-10000",
                                      a18 == 3 ~ "11000-19000",
                                      a18 == 4 ~ "20000-29000",
                                      a18 == 5 ~ "30000-39000",
                                      a18 == 6 ~ "40000-49000",
                                      a18 == 7 ~ "50000+")),
         income_1yr = as.factor(case_when(a18_rec1 == 0 ~ "$19,000",
                                      a18_rec1 == 1 ~ "$20,000-$49,000",
                                      a18_rec1 == 2 ~ "$50,000")),
         any_util = as.factor(case_when(any_util == 0 ~ "No", any_util == 1 ~ "Yes")),
         attempted_suicide = as.factor(case_when(g1c == 0 ~ "No", g1c == 1 ~ "Yes")),
         employment = as.factor(
           case_when(a13 == 1 ~ "Full time",
                     a13 == 2 ~ "Part time",
                     a13 == 3 ~ "Student",
                     a13 == 4 ~ "Unemployed",
                     a13 == 5 ~ "Ctrl_envir")),
         homeless = as.factor(case_when(homeless == 0 ~ "No", homeless == 1 ~ "Yes")),
         hs_grad = as.factor(case_when(hs_grad == 0 ~ "No", hs_grad == 1 ~ "Yes")),
         group = as.factor(case_when(group == 0 ~ "Control", group == 1 ~ "Clinic")),
         # linkstatus = as.factor(case_when(linkstatus == 0 ~ "Did not link to primary care", linkstatu
         alcohol = as.factor(case_when(alcohol == 0 ~ "Not First Drug", alcohol == 1 ~ "First Drug Alcol
         money_spent_on_alcohol = as.numeric(h16a),
         mh_index = as.numeric(mh),
         num_med_problems = as.numeric(d3),
         num_hospitilizations = as.numeric(d1),
         bothered_by_med = as.factor(case_when(d4 == 0 ~ "Not at all",
                                               d4 == 1 ~ "Slightly",
                                               d4 == 2 ~ "Moderately",
                                               d4 == 3 ~ "Considerably",
                                               d4 == 4 ~ "Extremely")),
         bothered = as.factor(case_when(d4_rec == 0 ~ "No",
                                        d4_rec == 1 ~ "Yes"))) %>%
  select(group, dayslink, linkstatus, yrs_education, gender, age, alcohol, alcq_30, marriage, employmen
```

```
## Warning: Problem with ‘mutate()‘ input ‘yrs_education‘.
## i NAs introduced by coercion
## i Input ‘yrs_education‘ is ‘as.numeric(a9)‘.

## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with ‘mutate()‘ input ‘alcq_30‘.
## i NAs introduced by coercion
## i Input ‘alcq_30‘ is ‘as.numeric(alcq_30)‘.

## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with ‘mutate()‘ input ‘money_spent_on_alcohol‘.
## i NAs introduced by coercion
## i Input ‘money_spent_on_alcohol‘ is ‘as.numeric(h16a)‘.

## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```
df %>%
  mutate(linkstatus = as.factor(case_when(linkstatus == 0 ~ "Did not link to primary care", linkstatus =
  select(group, linkstatus, dayslink, income, mh_index, attempted_suicide) %>%
  ggplot() + geom_point(aes(x=dayslink, y=mh_index, color=linkstatus)) + facet_grid(vars(attempted_suic
```

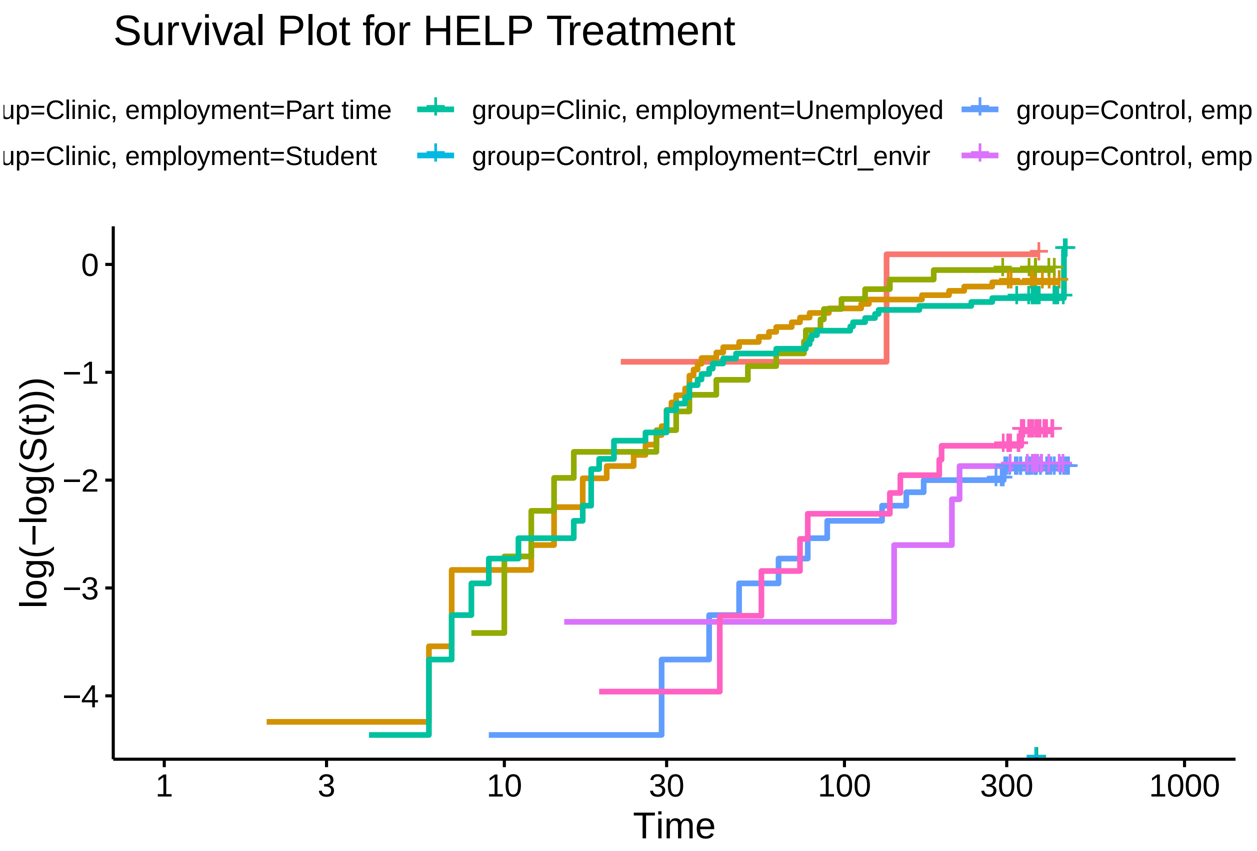


Cox PH Model

The explanatory variables we want to use are number of medical problems, gender, age, first drug alcohol, treatment, mental health index, and attempted suicide, and control.

First we will show the survival curves for just the treatment and control groups.

```
care_fit <- survfit(Surv(dayslink, linkstatus) ~ group + employment, data=df)
ggsurvplot(care_fit, conf.int = F, fun="cloglog") +
  ggtitle("Survival Plot for HELP Treatment")
```

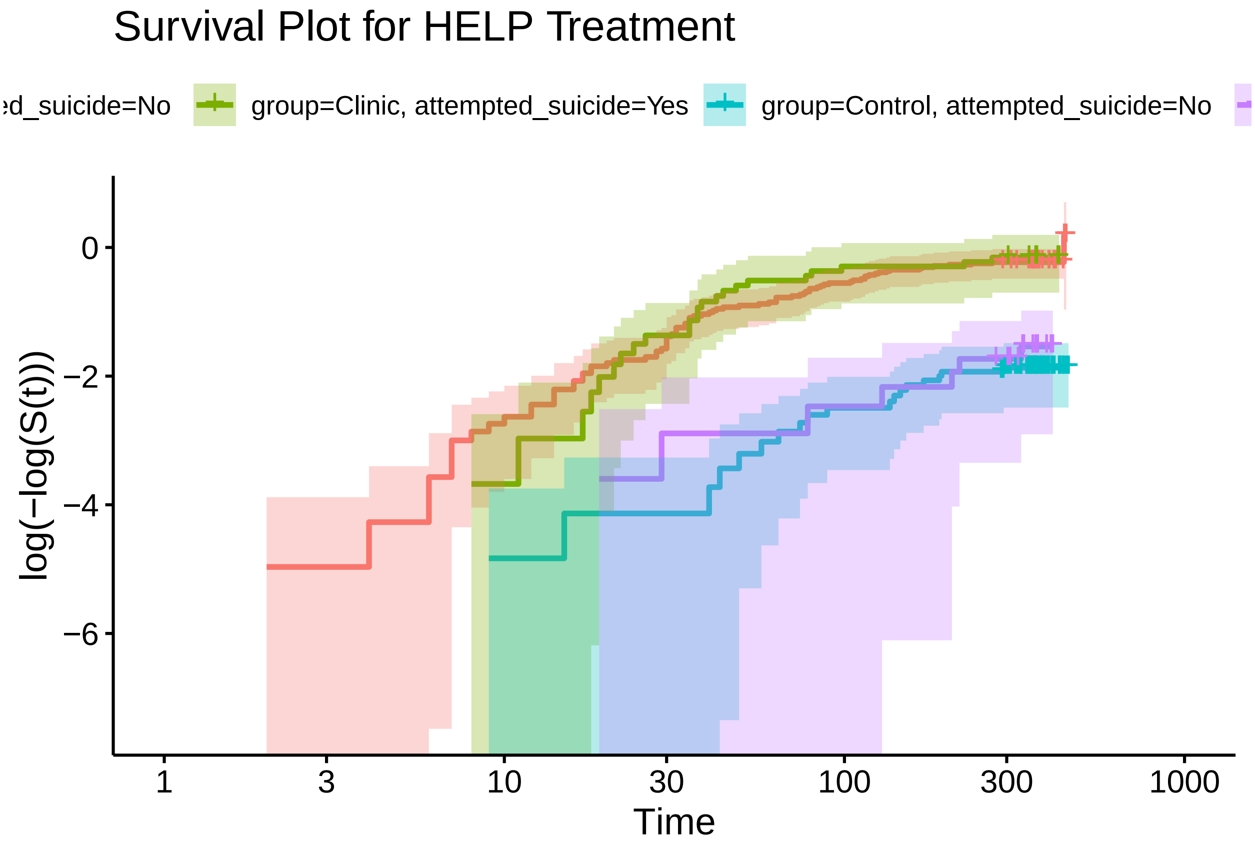


```r
coxph(Surv(dayslink, linkstatus) ~ group, data=df)
```

```
## Call:
## coxph(formula = Surv(dayslink, linkstatus) ~ group, data = df)
##
##                 coef exp(coef) se(coef)     z        p
## groupControl -1.6415    0.1937   0.2237 -7.34 2.14e-13
##
## Likelihood ratio test=69.59  on 1 df, p=< 2.2e-16
## n= 347, number of events= 128
```

The Cox-PH model for just the treatment group tells us that a unit increase in treatment to control groups (0 -> 1) changes the factor of survival by $e^{-1.6415} = 0.1936893$.

```r
care_fit <- survfit(Surv(dayslink, linkstatus) ~ group + attempted_suicide, data=df)
ggsurvplot(care_fit, conf.int = T, fun="cloglog") +
  ggtitle("Survival Plot for HELP Treatment")
```



```
coxph(Surv(dayslink, linkstatus) ~ group + attempted_suicide, data=df)
```

```
## Call:
## coxph(formula = Surv(dayslink, linkstatus) ~ group + attempted_suicide,
##     data = df)
##
##                          coef exp(coef) se(coef)      z        p
## groupControl          -1.6435    0.1933   0.2237 -7.347 2.02e-13
## attempted_suicideYes   0.1285    1.1371   0.2090  0.615    0.539
##
## Likelihood ratio test=69.96  on 2 df, p=6.436e-16
## n= 347, number of events= 128
```

```
coxph(Surv(dayslink, linkstatus) ~ group, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          69.6    7.30e-17         66.8   3.06e-16           53.9
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
base_loglik <- -683.2197
```

Next we'll compute the drop-in-deviance test to determine if attempted suicide should be included in our model.

$G = 2 * (logLik_{biggermodel} - logLik_{smallermodel})$

```
coxph(Surv(dayslink, linkstatus) ~ group + attempted_suicide, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          70.0    6.44e-16         67.1   2.65e-15           54.2
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-683.0351 - base_loglik)
1-pchisq(G, 1)
```

```
## [1] 0.5434407
```

The change in deviance is 0.3692, ($H_0 : \gamma = 0$), so with one degree of freedom the p-value is 0.5434407, which is greater than 0.05. We fail to reject the null hypothesis and do not need type in the model.

```
coxph(Surv(dayslink, linkstatus) ~ group + mh_index, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          69.8    7.04e-16         66.9   2.92e-15           54.0
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-683.1243 - base_loglik)
1-pchisq(G, 1)
```

```
## [1] 0.6622516
```

The change in deviance is 0.1908, ($H_0 : \gamma = 0$), so with one degree of freedom the p-value is 0.6622516, which is greater than 0.05. We fail to reject the null hypothesis and do not need type in the model.

```
coxph(Surv(dayslink, linkstatus) ~ group + gender, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          72.7    1.61e-16         69.8   7.09e-16            56.8
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-681.652 - base_loglik)
1-pchisq(G, 1)
```

```
## [1] 0.07660959
```

The change in deviance is 3.1354, ($H_0 : \gamma = 0$), so with one degree of freedom the p-value is 0.07660959, which is greater than 0.05. Note that the p-value is close to 0.05, which suggests that there could be little evidence. We fail to reject the null hypothesis and do not need type in the model.

```
coxph(Surv(dayslink, linkstatus) ~ group + alcohol, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          77.9    1.24e-17         74.9   5.38e-17            61.9
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-679.0874 - base_loglik)
1-pchisq(G, 1)
```

```
## [1] 0.004042557
```

The change in deviance is [1] 8.2646. ($H_0 : \gamma = 0$), so with one degree of freedom the p-value is 0.004042557, which is less than 0.05. We reject the null hypothesis that $\gamma = 0$ in favor of $H_a : \gamma \neq 0$ and should include first drink alcohol in the model.

```
coxph(Surv(dayslink, linkstatus) ~ group + num_med_problems, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          73.1    1.31e-16         70.2   5.58e-16            57.4
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-681.4424 - base_loglik)
1-pchisq(G, 1)
```

```
## [1] 0.05938062
```

We see that the p-value for the additive model for treatment groups and the number of medical problems is insignificant (0.05938062). Note that the additive model

```
coxph(Surv(dayslink, linkstatus) ~ group + alcohol, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          77.9    1.24e-17         74.9   5.38e-17           61.9
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
base_loglik_alc <- -679.0874 # update base loglikelihood for future drop-in-deviance tests
coxph(Surv(dayslink, linkstatus) ~ group + alcohol + num_med_problems, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          81.1    1.79e-17         77.6   9.91e-17           64.8
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-677.4701 - base_loglik_alc)
1-pchisq(G, 1)
```

```
## [1] 0.07209791
```

The p-value is 0.072. This does not suggest evidence to reject the null hypothesis that $\gamma = 0$.

```
coxph(Surv(dayslink, linkstatus) ~ group + alcohol + gender, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          79.6    3.66e-17         76.6   1.62e-16           63.6
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-678.1924 - base_loglik_alc)
1-pchisq(G, 1)
```

```
## [1] 0.1809262
```

```
coxph(Surv(dayslink, linkstatus) ~ group + alcohol + mh_index, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          78.3    7.26e-17         75.2   3.30e-16           62.2
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-678.8862 - base_loglik_alc)
1-pchisq(G, 1)
```

```
## [1] 0.5258524
```

```
coxph(Surv(dayslink, linkstatus) ~ group + alcohol + attempted_suicide, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          78.2    7.59e-17         75.1   3.40e-16           62.2
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-678.9307 - base_loglik_alc)
1-pchisq(G, 1)
```

```
## [1] 0.5756013
```

**** to use later ****

```
coxph(Surv(dayslink, linkstatus) ~ group + alcohol, data=df) %>% tidy()
```

```
## # A tibble: 2 x 5
##   term                   estimate std.error statistic  p.value
##   <chr>                     <dbl>     <dbl>     <dbl>    <dbl>
## 1 groupControl             -1.69      0.224     -7.54 4.58e-14
## 2 alcoholNot First Drug   -0.545     0.196     -2.78 5.41e- 3
```

Explanatory variables we want to use are gender, age, income, employment, homeless, and hs_grad.

```
coxph(Surv(dayslink, linkstatus) ~ group + age, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128           76.4    2.57e-17         73.0   1.40e-16           60.2
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-679.8128 - base_loglik)
1-pchisq(G, 1)
```

```
## [1] 0.009045607
```

0.009045607 < 0.05. We reject the null hypothesis and should include age.

```
coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##         n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##     <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128           81.8    1.24e-17         78.6   6.05e-17           65.7
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-677.0938 - -679.8128)
1-pchisq(G, 1)
```

```
## [1] 0.01970322
```

```
coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + employment, data=df) %>% glance()
```

```
## Warning in fitter(X, Y, istrat, offset, init, control, weights = weights, :
## Loglik converged before variable 6 ; coefficient may be infinite.
```

```
## # A tibble: 1 x 18
##         n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##     <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128           85.1    1.23e-15         81.4   7.03e-15           67.5
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-675.4454    - -677.0938)
1-pchisq(G, 1)
```

```
## [1] 0.06941499
```

```r
coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + homeless, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   347    128          82.0    6.63e-17         78.8   3.06e-16           65.8
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```r
G = 2*(-677.0244 -  -677.0938   )
1-pchisq(G, 1)
```

```
## [1] 0.7094769
```

```r
coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + income, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   336    126          89.1    2.44e-15         85.4   1.34e-14           70.6
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```r
G = 2*(-661.2074 -  -677.0938)
1-pchisq(G, 1)
```

```
## [1] 1.733028e-08
```

```r
coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + income + yrs_education, data=df) %>% glance(
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   333    125          95.4    4.59e-16         92.8   1.52e-15           77.3
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```r
G = 2*(-651.3298 - -661.2074)
1-pchisq(G, 1)
```

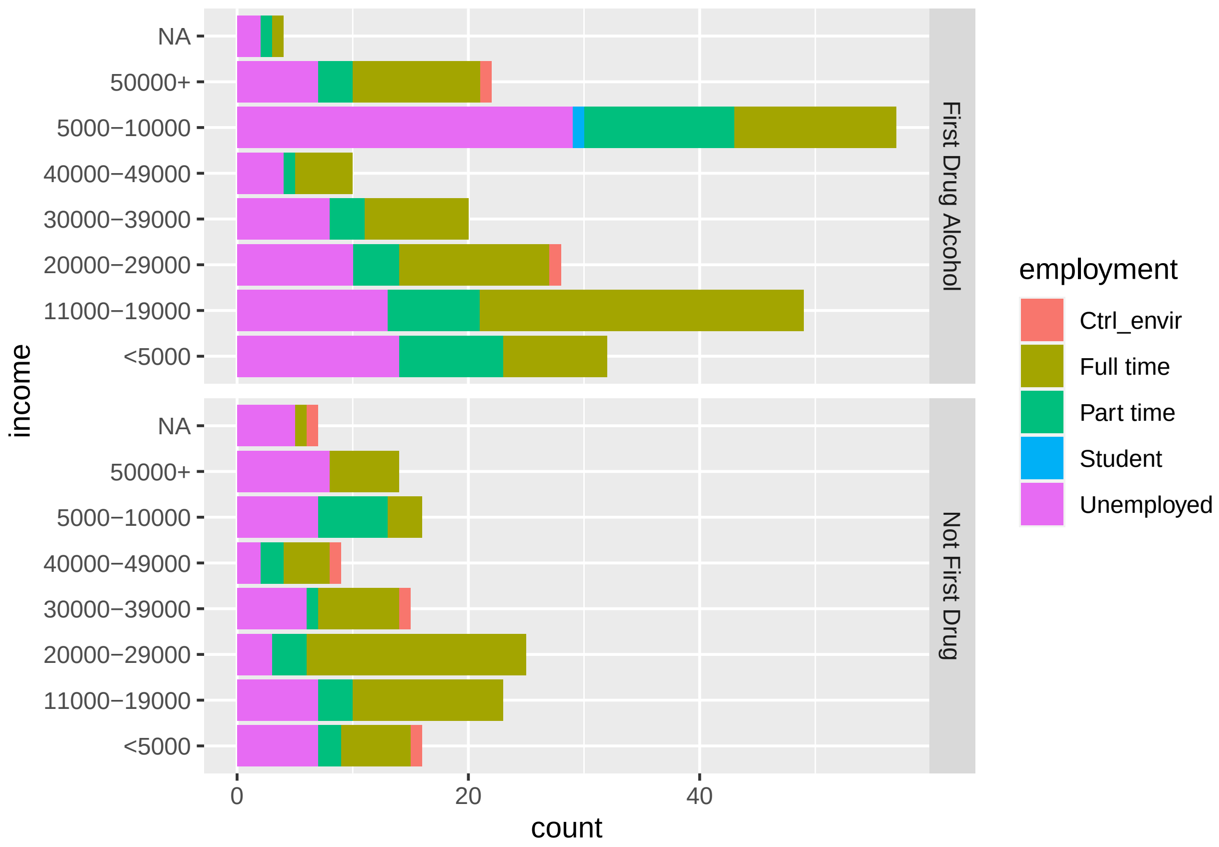
```
## [1] 8.802188e-06
```

```
coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + income + yrs_education, data=df)
```

```
## Call:
## coxph(formula = Surv(dayslink, linkstatus) ~ group + age + alcohol +
##     income + yrs_education, data = df)
##
##                           coef exp(coef) se(coef)      z        p
## groupControl          -1.84690   0.15772  0.23327 -7.917 2.43e-15
## age                    0.02591   1.02625  0.01209  2.143  0.03208
## alcoholNot First Drug -0.41864   0.65794  0.21025 -1.991  0.04647
## income11000-19000      0.32875   1.38923  0.30125  1.091  0.27515
## income20000-29000      0.24058   1.27199  0.32938  0.730  0.46514
## income30000-39000      0.10428   1.10991  0.43231  0.241  0.80939
## income40000-49000      0.70615   2.02617  0.40202  1.756  0.07900
## income5000-10000       0.15755   1.17064  0.30889  0.510  0.61003
## income50000+          -0.18031   0.83501  0.38408 -0.469  0.63875
## yrs_education         -0.13184   0.87648  0.04967 -2.654  0.00794
##
## Likelihood ratio test=95.37  on 10 df, p=4.586e-16
## n= 333, number of events= 125
##    (14 observations deleted due to missingness)
```

Above are the coefficients for the model. Note that the the p-values are income are all greater than 0.05. In fact, the p-value for income40000-49000 is smaller than the other income coefficients by a factor of 100. One reason behind the smaller p-value is that the amount of individuals with income between 40000-49000 were less than the other income groups.

```
df %>%
  select(income, employment, alcohol) %>%
  ggplot() + geom_bar(aes(x=income, fill=employment)) + facet_grid(vars(alcohol)) + coord_flip()
```



Looking the plot, we see a drop in the number of individuals for the 40000-49000 income group. So even though the drop-in-deviance test is significant, the outliers could be dragging the p-value down. Thus, we will not be proceeding with income.

```
coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data=df) %>% glance()
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   344    127          87.6    4.24e-18         84.8   1.70e-17           71.4
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education + hs_grad, data=df) %>% glance
```

```
## # A tibble: 1 x 18
##       n nevent statistic.log p.value.log statistic.sc p.value.sc statistic.wald
##   <int>  <dbl>         <dbl>       <dbl>        <dbl>      <dbl>          <dbl>
## 1   344    127          90.5    5.16e-18         88.2   1.57e-17           75.2
## # ... with 11 more variables: p.value.wald <dbl>, statistic.robust <dbl>,
## #   p.value.robust <dbl>, r.squared <dbl>, r.squared.max <dbl>,
## #   concordance <dbl>, std.error.concordance <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, nobs <int>
```

```
G = 2*(-665.9839 - -667.4519)
1-pchisq(G, 1)
```

```
## [1] 0.08662501
```

We will not be using the high school graduate binary variable.

After running the drop-in-deviance tests, we found age, alcohol, income, years in education, and high school graduate to be significant. Based on Occam's Razor, we are interested in picking a simpler model. Thus, we will proceed with Backward Selection.

Our final model:

```
coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data=df)
```

```
## Call:
## coxph(formula = Surv(dayslink, linkstatus) ~ group + age + alcohol + 
##     yrs_education, data = df)
## 
##                            coef exp(coef) se(coef)      z        p
## groupControl          -1.76952   0.17042  0.22661 -7.809 5.78e-15
## age                    0.02609   1.02643  0.01178  2.214   0.0268
## alcoholNot First Drug -0.42942   0.65088  0.20141 -2.132   0.0330
## yrs_education         -0.11815   0.88857  0.04722 -2.502   0.0124
## 
## Likelihood ratio test=87.61  on 4 df, p=< 2.2e-16
## n= 344, number of events= 127 
##    (3 observations deleted due to missingness)
```

```
# test for proportional hazards
cox.zph(coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data=df))
```

```
##                chisq df      p
## group         6.9850  1 0.0082
## age           0.0818  1 0.7749
## alcohol       1.1684  1 0.2797
## yrs_education 0.5564  1 0.4557
## GLOBAL        8.0528  4 0.0897
```

More formally,

$$h_i(t) = h_0 exp\{-1.76952 \cdot \text{Group} + 0.02609 \cdot \text{Age} - 0.42942 \cdot \text{Alcohol} - 0.11815 \cdot \text{YearsOfEducation}\}$$

The most drastic change in risk comes from a unit increase in group. That is, a unit increase in the binary encoding (0 -> 1), so treatment groups to control groups has a decrease in risk by a factor of $exp\{-1.76952\} = 0.1704148$