Cumulative Incidence Curve Regression

Zachary McCaw

2024-11-04

Purpose

This vignette illustrates performing regression of cumulative incidence probabilities via pseudo-values.

```
suppressPackageStartupMessages({
  library(dplyr)
  library(ggplot2)
  library(SurvUtils)
})
```

Pseudo-value regression

Simulate data

Time to the event of interest is simulated from an exponential distribution with rate parameter:

$$\lambda(A, X_1, X_2) = \lambda_0 \exp(A\beta_A + X_1\beta_1 + X_2\beta_2)$$

Here λ_0 is the base event rate, $A \in \{0, 1\}$ is the treatment arm, and (X_1, X_2) are covariates. The coefficients are set such that treatment $A = 1$ *decreases* the event rate by 50%, a one standard deviation increase in X_1 *increases* the event rate by 20%, and X_2 has no effect. Time to the event of interest is subject to a competing risk from death, with rate $\lambda_D = 0.25$, and independent censoring, with rate $\lambda_C = 0.25$.

```
set.seed(101)
# Generate covariate frame.
n <- 2 * 1e3
df_x <- data.frame(
  arm = rep(c(0, 1), each = 1e3),
  x1 = stats::rnorm(n),
  x2 = stats::rnorm(n)
)

# Center the scale the covariates.
df_x$x1 <- scale(df_x$x1)
df_x$x2 <- scale(df_x$x2)

# Simulate data.
df <- SurvUtils::GenCRData(
  base_death_rate = 0.25,
  base_event_rate = 1.0,
  beta_event = c(log(0.5), log(1.2), 0),
  censoring = 0.25,
  covariates = data.matrix(df_x)
)
df <- cbind(df, df_x)

# Tabulate censorings, events, and deaths by treatment arm.
df %>%
  dplyr::group_by(arm) %>%
  dplyr::summarise(
    n = dplyr::n(),
    n_censor = sum(status == 0),
    n_event = sum(status == 1),
    n_death = sum(status == 2)
  )
#> # A tibble: 2 x 5
#>     arm     n n_censor n_event n_death
#>   <dbl> <int>    <int>   <int>   <int>
#> 1     0  1000      169     656     175
#> 2     1  1000      261     498     241
```

Pseudo-values regression

Pseudo-values at a specified time point τ are generated with the `SurvUtils::GenPseudo` function. The pseudo-value is defined as:

$$\hat{\theta}_i(\tau) = n \cdot \hat{\theta}(\tau) - (n-1) \cdot \hat{\theta}_{(-i)}(\tau)$$

where $\hat{\theta}_i(\tau)$ is the pseudo-value for subject i at time τ, n is the sample size, $\hat{\theta}(\tau)$ is the value of the cumulative incidence curve (CIC) at time τ based on the full sample, and $\hat{\theta}_{(-i)}(\tau)$ is the CIC at time τ based on the jackknifed sample with subject i excluded. Pseudo-values take censoring into account during their construction, and once calculated, can be modeled in the same way as any continuous outcome.

Below, pseudo-values for the CIC at time $\tau = 2$ are modeled via linear regression. Consequently, the coefficients are interpreted as risk differences. For example, the coefficient on `arm` indicates that the treatment $(A = 1)$ is estimated to reduce the cumulative incidence of the event of interest at time $\tau = 2$ by 17.9%, holding (X_1, X_2) constant. Meanwhile, a standard deviation increase in X_1 is associated with a 6.7% increase in the cumulative incidence of the event of interest at time $\tau = 2$, holding (A, X_2) constant. As expected, A is associated with a significant reduction in the cumulative incidence, X_1 is associated with a comparatively smaller but significant increase, and X_2 has no significant effect.

```
# Generate pseudo-values for the cumulative incidence.
df <- SurvUtils::GenPseudo(df, tau = 2, type = "cic")

# Fit a linear pseudo-value regression.
fit <- stats::lm(pseudo ~ arm + x1 + x2, data = df)
summary(fit)
#>
#> Call:
#> stats::lm(formula = pseudo ~ arm + x1 + x2, data = df)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -1.0304 -0.5943  0.2392  0.4235  0.9211
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.71957    0.01666  43.196  < 2e-16 ***
#> arm         -0.17934    0.02356  -7.611 4.18e-14 ***
#> x1           0.06654    0.01178   5.648 1.86e-08 ***
#> x2          -0.01225    0.01179  -1.039    0.299
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.5267 on 1996 degrees of freedom
#> Multiple R-squared:  0.04289,    Adjusted R-squared:  0.04145
#> F-statistic: 29.82 on 3 and 1996 DF,  p-value: < 2.2e-16
```

Ground truth

The generative model was specified in terms of multiplicative effects of the treatment and covariates on the rate parameter for the event of interest, whereas our analysis estimates rate differences at a particular time point. Since the generative model for the data is known, we can determine the ground truth values for the rate differences by simulating a large data set in the absence of censoring, then fitting the pseudo-value regression at the same time point, as demonstrated below. Uncertainty in the ground truth parameter estimates can be made arbitrarily small by making the sample size sufficiently large.

```
set.seed(101)
# Generate covariate frame.
n <- 2 * 1e4
df_x <- data.frame(
  arm = rep(c(0, 1), each = 1e4),
  x1 = stats::rnorm(n),
  x2 = stats::rnorm(n)
)

# Center the scale the covariates.
df_x$x1 <- scale(df_x$x1)
df_x$x2 <- scale(df_x$x2)

# Simulate data.
df <- SurvUtils::GenCRData(
  base_death_rate = 0.25,
  base_event_rate = 1.0,
  beta_event = c(log(0.5), log(1.2), 0),
  censoring = 0.0,
  covariates = data.matrix(df_x)
)
df <- cbind(df, df_x)

# Generate pseudo-values for the cumulative incidence.
df <- SurvUtils::GenPseudo(df, tau = 2, type = "cic")

# Fit a linear pseudo-value regression.
fit <- stats::lm(pseudo ~ arm + x1 + x2, data = df)
summary(fit)
#>
#> Call:
#> stats::lm(formula = pseudo ~ arm + x1 + x2, data = df)
#>
```

```
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -0.8921 -0.5212  0.2487  0.3901  0.6841
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.725487   0.004707 154.143   <2e-16 ***
#> arm         -0.200775   0.006656 -30.164   <2e-16 ***
#> x1           0.052850   0.003328  15.880   <2e-16 ***
#> x2           0.004518   0.003328   1.358    0.175
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.4707 on 19996 degrees of freedom
#> Multiple R-squared:  0.05502,    Adjusted R-squared:  0.05488
#> F-statistic: 388.1 on 3 and 19996 DF,  p-value: < 2.2e-16
```