MA592 HW6

Zhiwei Liang

12/2/2021

```
dt = read.csv("HW6_2021_simAIEDRP_4500.csv")
#head(dt)
```

1. Specify a model for how the probability of treatment initiation, given that treatment was not initiated before, depends on injection drug use (this is a baseline covariate), month, and current CD4 count (this is a time-varying covariate).

$$for k > 6, logit Pr(A_k = 1 | \overline{A_{k-1}} = 0, injdrug, Y_k) = \theta_0 + \theta_1 injdrug + \theta_2 k + \theta_3 Y_k$$

$$for k = 6, logit Pr(A_6 = 1 |, injdrug, Y_6) = \theta_0 + \theta_1 injdrug + \theta_3 Y_6$$

2.Specify one overall model for the probability of treatment initiation, for all time points, including the baseline and later time points, without changing the assumptions of the model in 1. Hint: you may want to use interactions with 1k>6.

$$forall k, logit Pr(A_k = 1 | \overline{A_{k-1}} = 0, injdrug, Y_k) = \theta_0 + \theta_1 injdrug + I_{k>6}\theta_2 k + \theta_3 Y_k$$

3.Using the simulated data, fit the model you specified in 2. If you would like to check your answer to 2., you may also want to fit the models you specified in 1. and compare the results. If you are not sure, just fit any model for treatment initiation given the past, including at least the time points after month 6.

```
dt3 = dt %>% filter(month_firsttrt>=month) # treatment history = 0
fit3 = glm(treated~injdrug+monthgt6*month+cd4,data = dt3,family = binomial(link = "logit"))
summary(fit3)
```

```
##
## Call:
## glm(formula = treated ~ injdrug + monthgt6 * month + cd4, family = binomial(link = "logit"),
##     data = dt3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8341  -0.4425  -0.3413  -0.2937   2.7340
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.2801274  0.1311647 -25.008  < 2e-16 ***
## injdrug        -0.1393119  0.1468646  -0.949 0.342837
```

```
## monthgt6       -0.5078554  0.1237858  -4.103 4.08e-05 ***
## month           0.0990328  0.0046318  21.381  < 2e-16 ***
## cd4            -0.0005928  0.0001571  -3.773 0.000161 ***
## monthgt6:month         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11653  on 20497  degrees of freedom
## Residual deviance: 11039  on 20493  degrees of freedom
## AIC: 11049
##
## Number of Fisher Scoring iterations: 5
```

4.For a given treatment, do later CD4 counts depend on the variables which are pre- dictors of treatment initiation such as the current CD4 count? How do you find out? Choose a relatively simple solution first and see whether that already works to find a dependence.

Let's take the CD4 at time point k+1 as an example. I add a column in csv file represents the Y_{k+1}. And then I fit the model for the Y_{k+1} using the predictors of treatment initiation.

```
dt4 = dt %>% filter(month < 30)
fit4 = lm(CD4_kplus1 ~ injdrug+monthgt6*month+cd4,data = dt4)
summary(fit4)
```

```
##
## Call:
## lm(formula = CD4_kplus1 ~ injdrug + monthgt6 * month + cd4, data = dt4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -162.307  -17.986   -0.181   18.008  174.144
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    -1.047e+01  4.615e-01  -22.685   <2e-16 ***
## injdrug        -2.945e+00  3.164e-01   -9.308   <2e-16 ***
## monthgt6        5.504e-01  4.602e-01    1.196    0.232
## month           1.803e-01  1.347e-02   13.382   <2e-16 ***
## cd4             1.001e+00  3.436e-04 2913.092   <2e-16 ***
## monthgt6:month         NA         NA       NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.35 on 107995 degrees of freedom
## Multiple R-squared:  0.9898, Adjusted R-squared:  0.9898
## F-statistic: 2.609e+06 on 4 and 107995 DF,  p-value: < 2.2e-16
```

It seems like the coefficients of all the covariates, such as the current CD4, are significant. So the later CD4 counts do depend on the predictors.

5.Based on these data, do you think there is confounding by indication?

Through the former questions, the treatment A depends on outcome Y, so there's confounding by indication.

6.In this assignment, we focus on estimating how the mean of CD30 depends on treat- ment initiation. Using the simulated data and your model from 3., test H0: CD430 doesn't depend on treatment initiation. Hint: use ideas about testing described in the first 38 slides of the slide deck on Structural Nested Mean Models.

Add Y_30 to the model I fit in Q3.

$$logitPr(A_k = 1|\overline{A_{k-1}} = 0, injdrug, Y_k) = \theta_0 + \theta_1 injdrug + I_{k>6}\theta_2 k + \theta_3 Y_k + \alpha Y_{30}$$

The null hypothesis is equal to $\alpha = 0$. If reject $H_0 : \alpha = 0$, then treatment A affects the outcome Y.

```
dt6 = dt %>% filter(month_firsttrt>=month) # treatment history = 0
fit6 = glm(treated~injdrug+monthgt6*month+cd4+CD4_30,data = dt3,family = binomial(link = "logit"))
summary(fit6)
```

```
##
## Call:
## glm(formula = treated ~ injdrug + monthgt6 * month + cd4 + CD4_30,
##     family = binomial(link = "logit"), data = dt3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1924  -0.3986  -0.1938  -0.0784   3.3503
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.4015492  0.1790064 -30.175   <2e-16 ***
## injdrug        -0.1835351  0.1603299  -1.145   0.2523
## monthgt6        0.3288945  0.1470688   2.236   0.0253 *
## month           0.1694754  0.0055904  30.315   <2e-16 ***
## cd4            -0.0126826  0.0003472 -36.530   <2e-16 ***
## CD4_30          0.0122244  0.0003015  40.540   <2e-16 ***
## monthgt6:month         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11652.5  on 20497  degrees of freedom
## Residual deviance:  8589.3  on 20492  degrees of freedom
## AIC: 8601.3
##
## Number of Fisher Scoring iterations: 7
```

$\alpha = 0.0122$ and the p value is small enough, so we reject the null hypothesis. Instead, CD4_30 dose depend on treatment initiation.

7.Carry out a naive analysis, predicting the mean CD4 count in month 30 based on injection drug use, CD46, and the month of treatment initiation

```
fit7 = lm(CD4_30~injdrug+CD4_6+Tr_30,data = dt)
summary(fit7)
```

```
##
## Call:
## lm(formula = CD4_30 ~ injdrug + CD4_6 + Tr_30, data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -592.54  -94.01    1.29   93.32  517.06
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.385e+02  1.018e+00 -234.201   <2e-16 ***
## injdrug     -1.790e+00  1.556e+00   -1.151     0.25
## CD4_6        1.003e+00  1.854e-03  540.859   <2e-16 ***
## Tr_30        1.585e+01  5.600e-02  283.067   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 138.2 on 112496 degrees of freedom
## Multiple R-squared:  0.772,  Adjusted R-squared:  0.772
## F-statistic: 1.269e+05 on 3 and 112496 DF,  p-value: < 2.2e-16
```