Introduction to Causal Inference Exercise Assignment 1 - 2018

We use the SPSS pooled data of GSS 1983 and GSS 1991 for a causal analysis of the hours of watching TV (TVHOURS) with a dichotomized variable of church attendance as the treatment variable. The treatment variable ATDDMY is attending church at least once a month or more (ATTEND $\geq$4) versus less frequent attendance (ATTEND < 4) defined in the following SPSS syntax. We use categorical expression of age (AGE10C: 25-34, 35-44, 45-54, 55-64) , education (EDUC5: 0-11,12,13-15,16,17 and over), race (RACE_D: African Americans versus other races), employment status (WD: not employed versus employed, defined in the following syntax), and the interaction effect of WD and ages 35-54 (INT, defined in the following syntax) predicting ATDDMY and TVHOURS. The sample selection and the definition of control variables are as follows.

SPSS Syntax:

```
GET FILE=GSS83_91.SAV.
MISSING VALUES TVHOURS(-1,98,88), WRKSTAT(0,9), ATTEND(9).
SELECT IF (WRKSTAT NE 0 AND WRKSTAT LT 8).
SELECT IF (AGE GE 25).
SELECT IF (AGE LT 65).
SELECT IF (TVHOURS NE -1 AND TVHOURS LT 98).
SELECT IF (ATTEND LT 9).
SELECT IF (EDUC LT 97).
COMPUTE WD=range(WRKSTAT,3,8).
COMPUTE RACE_D=range(race,2,2).
COMPUTE ATDDMY=range(ATTEND,4,8).
COMPUTE EDUC5=EDUC.
RECODE EDUC5 (0 thru 11=1)(12=2)(13 thru 15=3)(16=4)(17 thru 20=5).
COMPUTE AGE10C=TRUNC((AGE-5)/10).
COMPUTE INT=WD*RANGE(AGE10C,3,4).

LOGISTIC REGRESSION VARIABLES ATDDMY
  /METHOD=ENTER EDUC5 RACE_D AGE10C WD INT
  /CATEGORICAL EDUC5 AGE10C
   /CONTRAST (EDUC5)=Indicator(2)
   /CONTRAST(AGE10C)=indicator(1)
    /SAVE PRED
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).

RENAME VAR (PRE_1=PROPEN).
COMPUTE LGT_P=LN(PROPEN/(1.0-PROPEN)).
COMPUTE LGT_P10=TRUNC(LGT_P*10+0.5).GET FILE=GSS83_91.SAV.
```

The rest is omitted

Answer the following questions.

1. Run the logistic regression to predict the treatment variable ATDDMY including AGE10C, EDUC5, RACE_D, WD, INT as covariates . Treat AGE10C and EDUC5 as categorical variables by using the first category (25-34) as the baseline category for age, and the second category (high school graduation) as the baseline category for education.

2. Describe the effects of (1) race dummy variable, (2) the effect of employment status for (a) ages 35-54, and (b) other ages on the odds of attending churches at least once a month.

3. Using the cross-classification of LGT_P10 (which is ten times of the logit of the propensity score rounded to an integer value) and ATDDMY, identify (a) the area of common support by the values of the LGT-P10, and (b) the number of sample subjects outside of the common area of support.

4. Define 0 strata as follows using the LGT_P10.
(1) [-5,-3], (2) [-2,-1], (3) 0, (4) 1, (5) 2, (6) 3, (7) [4,6], (8) [7,9], (9)[10,14]

5. Calculate the mean of the logit of propensity score (LOGIT_P) by strata for each category of the treatment variable.

6. Test the significance of (1) the difference in the logit between the treatment group and the control group for each stratum, and (2) estimate the average difference in the logit weighted by the proportion of samples in each stratum, and its standard error. Confirm that no significant differences in the logit exist.

7. Test the conditional independence between ATDDMY and each of the four variables (EDUC5, RACE_D, AGE10C, and WD) controlling for strata.

8. Calculate the mean of the outcome variable (TVHOURS) for each stratum.

9. Calculate the estimate for the ATE and its standard error, and conclude whether church attendance significantly affects the hours of watching TV or not.

10. Using the estimate of the propensity score, (1) create weights for ATDDMY=1 and those for ATDDMY=0 by the IPT formula for the ATE.

11. Check whether the sum of weighed frequencies for each of ATDDMY's states is close to the number of samples for each state of ATDDMY. What can you tell from this diagnosis?

12. Adjust the IPT weights to have the average weight to be 1.0 for each state of the treatment variable.

13. Calculate the correlation between ATDDMY and the logit of the propensity score using the IPT weights. Conform that correlation is nearly 0.

14. Test the statistical independence between each of the four covariates and ATTDMY with the IPT-weighted data.

15. Conduct the T-test for the difference in the weighted mean of TVHOURS between two states of the treatment variable (ATDDMY), give estimate for the treatment effect, its standard error, and conclude. Assume a distinct variance of Y for each state of the treatment variable.

16. Apply the following two linear regression models using TVHOURS as the dependent variables. (1) regression model with ATDDMY as the only explanatory variable with the IPT weights, and (2) regression model with ATDDMY and the control variables (use dummy variables for the education and age variables) as the explanatory variables with the IPT weights.

17. For the results from (16), answer the following.

(a) Confirm that the treatment effect in the model of (1) is the same as the one we obtained in (15), but the standard error slightly differs. Why does it differ?
(b) What is the purpose of applying model (2)? Why we should expect an estimate of the treatment effect that is very close to the result of model (1).