**2023 Causal Inference Workshop: Stata, R, and Python Sessions**

**Day 1. Assess balance between no\_insurance\_wave\_1 and control groups**

Underlying Paper: Black, Espin-Sanchez, French, and Litvak, *The Long-term Effect of Health Insurance on Near-Elderly Health and Mortality*, 3 **American Journal of Health Economics** 281-311 (2017)

**Answer Sheet**

1. **Question**: Create propensity scores for all insured versus uninsured.

The treatment used in the paper is **un**insurance at wave 1 in 1992. Variable name = no\_insurance\_wave\_1.

Logit: All insured (1 =uninsured) regressed on quadratic in age, gender, race, age-gender-race interactions, comorbidities, quintiles of log income and earnings, comorbidities, CESD depression score, self-reported health status, mobility limit, years of schooling, household size, and employment status dummies.

Extract predicted values (predict pscore, pr) and calculate IPW.

Initial Sample Size is 8203 for insured, and 1440 for uninsured.

**STATA Code for Step 1**:

**\*Create local variables for regressions**

local two\_power\_age = "age c.age#c.age" //quadratic in age

local comorbidities = "i.diabetes\_wave\_1 i.cancer\_wave\_1 i.high\_blood\_pressure\_wave\_1 i.lung\_disease\_wave\_1 i.heart\_disease\_wave\_1 i.stroke\_wave\_1 i.psychiatric\_disease\_wave\_1 i.arthritis\_wave\_1 i.ulcer\_wave\_1 i.CESD\_wave\_1”

local money = "i.hh\_logincome\_quintile\_wave\_1 i.hh\_earning\_quintile\_wave\_1" //quintiles of logincome and earnings

local labor = "i.not\_in\_laborforce\_wave\_1 i.partly\_retired\_wave\_1 i.fully\_retired\_wave\_1 i.unemployed\_wave\_1 i.employed\_pt\_wave\_1 i.employed\_ft\_wave\_1 i.veteran\_wave\_1"

**\*Tabulate no\_insurance\_wave\_1**

tab no\_insurance\_wave\_1

**\*Use relative age values to avoid scale problems**

replace age=age-51 //age now starts at 0 and ends at 10

**\*Logit regression**

logit no\_insurance\_wave\_1 `two\_power\_age' i.gender i.race `money' `labor' `comorbidities' i.gender#i.race i.gend er#c.(`two\_power\_age') i.race#c.(`two\_power\_age'), robust

**\*Predict propensity score**

predict pscore, pr //propensity to be uninsured

**\*Generate inverse propensity weights**

gen ipw=1/pscore if no\_insurance\_wave\_1==1 //weights for those who are not insured

replace ipw=1/(1-pscore) if no\_insurance\_wave\_1==0 //weights for those who are insured

1. **Question**: Replicate Black et al. Table 1 for uninsured versus all insured (ok to do this only for selected covariates)

The table below reports summary statistics table before and after IPW weighting for selected variables

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Covariate** | **Unweighted** | | **IPW** | | **t-test on Difference** | | **Normalized Difference** | |
|  | Uninsured | Insured | Uninsured | Insured | Unweighted | IPW | Unweighted | IPW |
| Age | 55.52 | 55.85 | 55.89 | 55.81 | -3.716 | 0.875 | -0.106 | 0.025 |
| Gender | 44.03 | 47.78 | 46.99 | 47.17 | -2.639 | -0.127 | -0.075 | -0.004 |
| White | 52.29 | 75.11 | 72.27 | 71.62 | -16.290 | 0.506 | -0.488 | 0.014 |
| Black | 21.81 | 16.25 | 16.17 | 17.05 | 4.780 | -0.835 | 0.142 | -0.024 |
| Hispanic | 22.57 | 6.78 | 9.14 | 9.16 | 13.896 | -0.022 | 0.458 | -0.001 |
| Other | 3.33 | 1.87 | 2.42 | 2.17 | 2.959 | 0.577 | 0.092 | 0.017 |
| Log(Income) | 9.17 | 10.40 | 10.19 | 10.26 | -19.426 | -1.484 | -0.653 | -0.045 |
| Diabetes | 11.11 | 9.08 | 9.19 | 9.54 | 2.287 | -0.422 | 0.067 | -0.012 |
| Cancer | 3.61 | 4.94 | 5.04 | 4.75 | -2.425 | 0.468 | -0.066 | 0.013 |
| Employed FT | 36.81 | 58.42 | 52.75 | 55.11 | -15.628 | -1.660 | -0.443 | -0.047 |
| Veteran | 17.85 | 27.05 | 26.67 | 25.66 | -8.201 | 0.805 | -0.222 | 0.023 |

**STATA Code to generate the summary statistics table above**

**\*Generate single dummies for race**

tab race, gen(race\_)

**\*Bring back age to its initial value**

replace age=age+51

**\*Generate weight equal 1 for unweighted (gives a weight of 1 to all observations)**

\* Not necessary, but allows a single loop and makes it easier to generate table in Excel

gen unwgtd=1

**\*Calculate t-test and normalized difference for control and treatment for selected variables**

local weights = "unwgtd ipw"

local vars = "age gender race\_1 race\_2 race\_3 race\_4 log\_income\_wave\_1 diabetes\_wave\_1 cancer\_wave\_1 employed\_ft\_wave\_1 veteran\_wave\_1"

**\*Loop to build summary stats table**

foreach y of local weights { //loop over unweighted and IPW

foreach j of local vars { //loop over selected vars

**\*Calculate the mean and standard deviation for the insured, and save to local variables**

quietly sum `j' if no\_insurance\_wave\_1==0 [aw=`y']

local mean\_insured = r(mean) //save mean

local sd\_insured = r(sd) //save sd

local sample\_insured = r(N) //save sample size

**\*Calculate mean and standard deviation for the uninsured, and save to local variables**

quietly sum `j' if no\_insurance\_wave\_1==1 [aw=`y']

local mean\_uninsured = r(mean) //save mean

local sd\_uninsured = r(sd) //save sd

local sample\_uninsured = r(N) //save sample size

**\* This code computes the normalized difference manually, and then closes the loops**

gen ttest\_`j'\_`y' = (`mean\_uninsured' - `mean\_insured')/(((`sd\_uninsured'^2/`sample\_uninsured' + `sd\_insured'^2/`sample\_insured'))^(1/2))

gen ndiff\_`j'\_`y' = (`mean\_uninsured' - `mean\_insured')/(((`sd\_uninsured'^2 + `sd\_insured'^2)/2) ^(1/2))

}

}

**\* Alternative using covbal.ado:**

ssc install covbal //to install the covbal package

\*Reference: (Linden, Ariel (2016). covbal: Stata module for generating covariate balance statistics. <https://ideas.repec.org/c/boc/bocode/s458188.html>)

\* covbal does not generate t-values, and does not allow direct export to Excel

local vars = "no\_insurance\_wave\_1 age gender race\_1 race\_2 race\_3 race\_4 log\_income\_wave\_1 diabetes\_wave\_1 cancer\_wave\_1 employed\_ft\_wave\_1 veteran\_wave\_1"

\* the first variable in the list must be the treatment

covbal `vars', wt(ipw) format(%9.3f) saving(“$path\summary\_stats\_using\_covbal\_ipw.dta”)

covbal `vars', wt(unwgtd) format(%9.3f) saving(“$path\summary\_stats\_using\_covbal\_unweighted.dta”)

**\*Generate summary stats**

foreach y of local weights{

preserve

\*Collapse mean

collapse (mean) age gender race\_\* log\_income\_wave\_1 diabetes\_wave\_1 cancer\_wave\_1 employed\_ft\_wave\_1 veteran\_wave\_1 ttest\_\*`y' ndiff\_\*`y' [aw=`y'], by(no\_insurance\_wave\_1) //generate means of each variable using different weights based on whether an individual is uninsured or insured.

\*Cleaning

replace gender=gender\*100

replace race\_1=race\_1\*100

replace race\_2=race\_2\*100

replace race\_3=race\_3\*100

replace race\_4=race\_4\*100

replace diabetes\_wave\_1=diabetes\_wave\_1\*100

replace cancer\_wave\_1=cancer\_wave\_1\*100

replace employed\_ft\_wave\_1=employed\_ft\_wave\_1\*100

replace veteran\_wave\_1=veteran\_wave\_1\*100

\*Rename

rename no\_insurance\_wave\_1 insured

\*Export summary stats to excel

export excel "$desktop\summary\_stats\_`date'", sheet("`y' no trim", replace) firstrow(variables)

restore

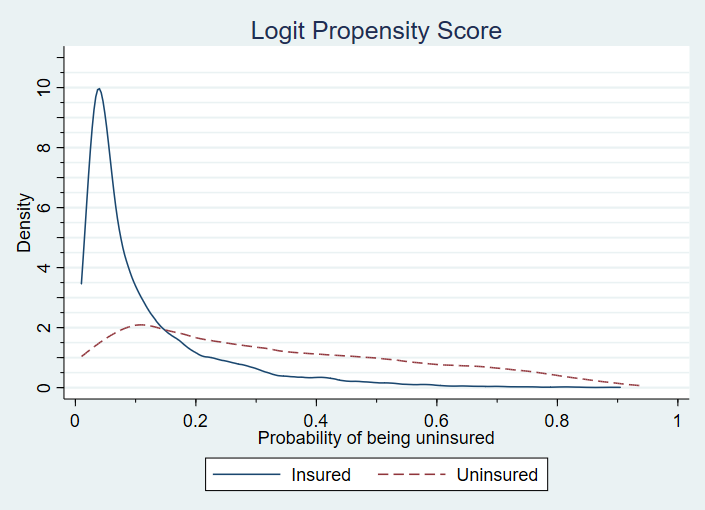
}

**\*Drop ttest values as we need to calculate them later again after trim.**

drop ttest\_\* ndiff\*

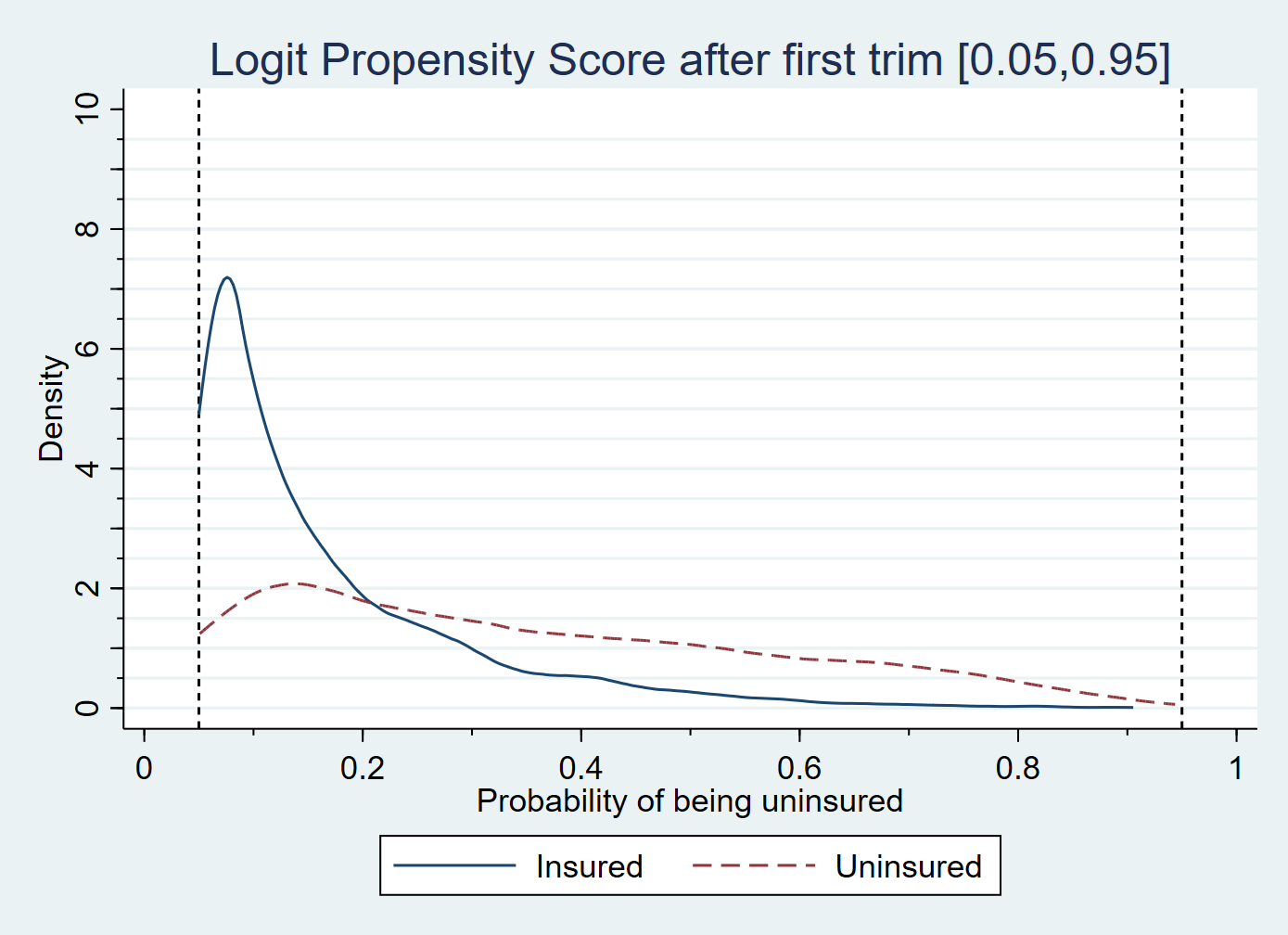
1. **Question**: Create kernel density plots, similar to Black et al. Figure 1, but without trimming to common support

First step Propensity Score using Kernel Density command in Stata



1. **Question**: Now trim to *p* between 0.05 and 0.95 (or to common support if narrower). What is remaining sample size for each comparison?

Remaining sample size after first trim: 5185 insured, 1336 uninsured



**STATA code for Steps 3 and 4:**

**\*Use kernel density for propensities**

twoway (kdensity pscore if no\_insurance\_wave\_1==1, lpattern(dash) lcol(red)) (kdensity pscore if no\_insurance\_wave\_1==0, lcol(blue)), /\*

\*/ legend(label( 1 "Uninsured") label( 2 "Insured")) /\*

\*/ xtitle("Probability of being uninsured") /\*

\*/ ytitle("Density") /\*

\*/ title("Logit Propensity Score") /\*

\*/ ytick(0(1)11, grid) ymtick(0(0.5)11, grid) ylab(0 "0" 2 "2" 4 "4" 6 "6" 8 "8" 10 "10")/\*

\*/ xmtick(0(0.1)1) xlabel(0 "0" 0.2 "0.2" 0.4 "0.4" 0.6 "0.6" 0.8 "0.8" 1 "1")

1. **Question**: Re-estimate propensity scores; create new kernel density plots

Remaining sample after recomputing propensity score and second trim: 5174 insured, 1333 uninsured.

