

* Stata Analysis Examples Replication for ASDA 3rd Edition
* Berglund Fall 2024

*** CHAPTER 11

* NOTE: all cases with KFINR = 1 and KWGTR > 0

use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrswide.dta", clear

```
gen ln_inc06 = ln(H8ITOT + 1)
gen ln_inc08 = ln(H9ITOT + 1)
gen ln_inc10 = ln(H10ITOT + 1)
gen ln_inc12 = ln(H11ITOT + 1)
```

```
hist ln_inc06
hist ln_inc08
hist ln_inc10
hist ln_inc12
```

* Table 11.2

```
* Complete case analysis for 2008 log-income.
svyset secu [pweight = kwgtr], strata(stratum)
svy: mean ln_inc08 if kfinr == 1
matrix list r(table)
matrix a = r(table)
di exp(a[1,1])
di exp(a[5,1])
di exp(a[6,1])
```

* Weight adjustment approach for 2008 log-income.

```
* Compute response indicator for 2008.
gen resp08 = 1 if ln_inc08 != .
replace resp08 = 0 if ln_inc08 == .
tab resp08, miss
```

```
* Modal imputation of missing covariate values.
replace selfrhealth_06 = 3 if selfrhealth_06 == .
replace marcat_06 = 2 if marcat_06 == .
replace diabetes_06 = 0 if diabetes_06 == .
replace arthritis_06 = 1 if arthritis_06 == .
replace racecat = 2 if racecat == .
replace edcat = 2 if edcat == .
```

```
* Response propensity model, with predictions.
svyset secu [pweight = kwgtr], strata(stratum)
svy: logit resp08 ln_inc06 i.selfrhealth_06 age_06 ///
    i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat, p
```

```
* Form deciles of response propensity, and compute means
* within each decile for weight adjustment.
xtile dec = phat, nq(10)
egen mean_phat = mean(phat), by(dec)
gen adj_kwgtr = kwgtr * (1 / mean_phat)
```

```
* Estimation using adjusted weights.
svyset secu [pweight = adj_kwgtr], strata(stratum)
svy: mean ln_inc08
matrix a = r(table)
di exp(a[1,1])
di exp(a[5,1])
di exp(a[6,1])
```

* MI Approach

```
* Create deciles of the 2006 survey weights.
xtile wgt_dec = kwgtr, nq(10)
```

```

* Set data structure and register both complete "regular"
* variables and variables with missing data to be imputed.
mi set flong
mi register imputed ln_inc08
mi register regular ln_inc06 selfrhealth_06 age_06 ///
    marcat_06 diabetes_06 arthritis_06 racecat edcat ///
    stratum wgt_dec
mi describe

* Implement regression imputation after setting a seed.
set seed 41279

mi impute chained (regress) ln_inc08 = ln_inc06 ///
    i.selfrhealth_06 age_06 i.marcat_06 diabetes_06 ///
    arthritis_06 i.racecat i.edcat i.wgt_dec i.stratum, ///
    noisily augment add(5) burnin(5)

* Set complex sampling features in MI framework.
mi svyset secu [pweight = kwgtr], strata(stratum)

* Multiple Imputation estimation.
mi estimate, vartable: svy: mean ln_inc08
matrix list r(table)
matrix a = r(table)
di exp(a[1,1])
di exp(a[5,1])
di exp(a[6,1])

* Selection model approach (need original data again)

* NOTE: all cases with KFINR = 1 and KWGTR > 0

use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrswide.dta", clear

gen ln_inc06 = ln(H8ITOT + 1)
gen ln_inc08 = ln(H9ITOT + 1)
gen ln_inc10 = ln(H10ITOT + 1)
gen ln_inc12 = ln(H11ITOT + 1)

* Compute response indicator for 2008.
gen resp08 = 1 if ln_inc08 != .
replace resp08 = 0 if ln_inc08 == .

svyset secu [pweight = kwgtr], strata(stratum)

svy: heckman ln_inc08 ln_inc06 selfrhealth_06 age_06 ///
i.marcat_06 i.edcat, ///
select(resp08 = diabetes_06 arthritis_06 ///
selfrhealth_06 age_06 i.marcat_06)

* Example 11.3.2 Estimation of change from 2006 to 2010
* Table 11.3
* Complete case analysis: mean change in income, 2006 to 2010.

gen incdiff_06_10 = H10ITOT - H8ITOT
svyset secu [pweight = kwgtr], strata(stratum)
svy: mean incdiff_06_10

* Weighting Adjustment

* Compute response indicator for 2010.
gen resp10 = 1 if ln_inc10 != .
replace resp10 = 0 if ln_inc10 == .
tab resp10, miss

* Modal imputation of missing covariate values.
replace selfrhealth_06 = 3 if selfrhealth_06 == .
replace marcat_06 = 2 if marcat_06 == .

```

```

replace diabetes_06 = 0 if diabetes_06 == .
replace arthritis_06 = 1 if arthritis_06 == .
replace racecat = 2 if racecat == .
replace edcat = 2 if edcat == .

* Response propensity model.
svyset secu [pweight = kwgtr], strata(stratum)
svy: logit resp10 ln_inc06 i.selfrhealth_06 age_06 ///
    i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat, p

* Form deciles of response propensity, and compute means
* within each decile for weight adjustment.
xtile dec = phat, nq(10)
egen mean_phat = mean(phat), by(dec)
gen adj_kwgtr = kwgtr * (1 / mean_phat)

* Estimation of mean change using adjusted weights.
svyset secu [pweight = adj_kwgtr], strata(stratum)
svy: mean incdiff_06_10

* Multiple imputation of 2010 log-income.

* Create deciles of the 2006 sampling weights.
xtile wgt_dec = kwgtr, nq(10)

* Set data structure and register both complete "regular"
* variables and variables with missing data to be imputed.
mi set flong
mi register imputed ln_inc10
mi register regular ln_inc06 selfrhealth_06 age_06 ///
    marcat_06 diabetes_06 arthritis_06 racecat edcat ///
    stratum wgt_dec
mi describe

* Implement regression imputation
set seed 41279

mi impute chained (regress) ln_inc10 = ln_inc06 i.selfrhealth_06 age_06 i.marcat_06 diabetes_06 ///
    arthritis_06 i.racecat i.edcat ///
    i.wgt_dec i.stratum, ///
    noisily augment add(5) burnin(5)

* Compute bounded change scores in each imputed data set.
sum incdiff_06_10 if _mi_m == 0
sum ln_inc10 if _mi_m == 0
replace ln_inc10=14.92 if ln_inc10 > 14.92 & ln_inc10 != .
gen new_chg0610 = exp(ln_inc10) - exp(ln_inc06)
replace new_chg0610 = -12300000 if new_chg0610 < -12300000
replace new_chg0610 = 2062968 if new_chg0610 > 2062968 & new_chg0610 != .

* Set complex sampling features for MI analysis.
mi svyset secu [pweight = kwgtr], strata(stratum)

* Multiple Imputation estimation.
mi estimate, vartable noisily: svy: mean new_chg0610

* Selection model approach (need original data again)
* NOTE: all cases with KFINR = 1 and KWGTR > 0

use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrswide.dta", clear

gen ln_inc06 = ln(H8ITOT + 1)
gen ln_inc08 = ln(H9ITOT + 1)
gen ln_inc10 = ln(H10ITOT + 1)
gen ln_inc12 = ln(H11ITOT + 1)

* Compute response indicator for 2008.
gen resp10 = 1 if ln_inc10 != .

```

```

replace resp10 = 0 if ln_inc10 == .

svyset secu [pweight = kwgtr], strata(stratum)

svy: heckman ln_inc10 ln_inc06 selfrhealth_06 age_06 ///
i.marcat_06 i.edcat, ///
select(resp10 = diabetes_06 arthritis_06 ///
selfrhealth_06 age_06 i.marcat_06)

* Calibration

* NOTE: all cases with KFINR = 1 and KWGTR > 0

use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrswide.dta", clear

* Modal imputation of missing covariate values.
replace racecat = 2 if racecat == .
replace edcat = 2 if edcat == .

* Compute sums of 2006 weights in cross-classes
* defined by sex, race, and education.
egen cal_class = group(racecat edcat gender)
collapse (sum) kwgtr, by(cal_class)
rename kwgtr popsize

save "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\cal_pop_sizes.dta", replace

* Repeat process for cases with complete data.

use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrswide.dta", clear

* Modal imputation of missing covariate values.
replace racecat = 2 if racecat == .
replace edcat = 2 if edcat == .

gen ln_inc10 = ln(H10ITOT + 1)

* Compute response indicator for 2010.
gen resp10 = 1 if ln_inc10 != .
replace resp10 = 0 if ln_inc10 == .
tab resp10, miss

* Only keep cases with complete data.
keep if resp10 == 1

* Compute sums of 2006 weights in cross-classes
* defined by sex, race, and education.
egen cal_class = group(racecat edcat gender)
collapse (sum) kwgtr, by(cal_class)
rename kwgtr sumrespwgts

save "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\cal_resp_sizes.dta", replace

* Merge the two data sets of estimated population sizes.

merge 1:1 cal_class using "P:\ASDA3\Data Sets for Analysis Examples and Stata R
Code\cal_pop_sizes.dta"

* Compute calibration adjustments for each cross-class.
gen cal_adj = popsize / sumrespwgts
drop _merge

save "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\cal_resp_sizes.dta", replace

* Open original data set and merge calibration adjustments.
use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrswide.dta", clear

* Modal imputation of missing covariate values.
replace racecat = 2 if racecat == .
replace edcat = 2 if edcat == .

```

```

egen cal_class = group(racecat edcat gender)
sort cal_class
gen ln_inc06 = ln(H8ITOT + 1)
gen ln_inc10 = ln(H10ITOT + 1)

merge m:1 cal_class using "P:\ASDA3\Data Sets for Analysis Examples and Stata R
Code\cal_resp_sizes.dta"

* Compute response indicator for 2010.
gen resp10 = 1 if ln_inc10 != .
replace resp10 = 0 if ln_inc10 == .
tab resp10, miss

* Compute calibrated weights for cases with complete data.
gen kwgtr_cal = kwgtr * cal_adj if resp10 == 1

* Verify that sums of calibrated weights for cases
* with complete data are equal to sums of base weights
* for full sample.
total kwgtr, over(cal_class)
total kwgtr_cal if resp10 == 1, over(cal_class)

* Estimate mean change using complete cases.
gen incdiff_06_10 = H10ITOT - H8ITOT

svyset secu [pweight = kwgtr_cal], strata(stratum) ///
poststrata(cal_class) postweight(popsize)

svy: mean incdiff_06_10
* End of Table 11.3 Numbers

*****
* Second-stage example

use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrssub0608.dta", clear

* Ignore individuals as second-stage clusters.
svyset secu [pweight = weight], strata(stratum)
svy: mean income, over(wave)
lincom c.income@2008.wave - c.income@2006.wave

* Account for individuals (HHIDPN) as second-stage clusters,
* incorporating small FPCs.
gen samprate = 0.01
svyset secu [pweight = weight], strata(stratum) ///
fpc(samprate) || hhidpn, fpc(samprate)
svy: mean income, over(wave)
lincom c.income@2008.wave - c.income@2006.wave

* Weighted multilevel modeling

use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrswide.dta", clear

gen ln_inc1 = ln(H8ITOT + 1)
gen ln_inc2 = ln(H9ITOT + 1)
gen ln_inc3 = ln(H10ITOT + 1)
gen ln_inc4 = ln(H11ITOT + 1)

* Initial wave-specific survey weights.
gen wgt1 = kwgtr
gen wgt2 = lwgtr
gen wgt3 = mwgtr
gen wgt4 = nwgtr

* Baseline sampling weight for respondents from 2006.
gen basewgt = kwgtr

```

```

* Keep variables of interest.
keep hhid pn gender ln_inc1-ln_inc4 wgt1-wgt4 ///
secu stratum basewgt

* Reshape data set from wide to vertical format.
reshape long ln_inc wgt, i(hhid pn) j(year)

* Drop missing income values (wave nonresponse).
drop if ln_inc == .

* Form a unique ID variable.
egen newid = concat(hhid pn)
destring newid, gen(newid_num)

* Compute conditional Level 1 weights
* (Skinner and Holmes 2003), and rescale weights.
gen levellwgt = wgt / basewgt
gen level2wgt = basewgt

* "Method 2" rescaling (normalizing weights).
egen sumw = sum(levellwgt), by(newid)
egen nj = count(ln_inc), by(newid)
gen levellwgt_r = levellwgt * nj/sumw

gen yrssince06 = 0 if year == 1
replace yrssince06 = 2 if year == 2
replace yrssince06 = 4 if year == 3
replace yrssince06 = 6 if year == 4

* Plot data for small subsample of subjects.
xtset newid_num year
xtline ln_inc if newid_num <= 10200000, overlay ///
legend(off) ytitle(Log-Transformed Annual Income) ///
tttitle(HRS Wave (1 = 2006))

* Plot means across waves for males and females.
collapse (mean) ln_inc, by(gender year)
xtset gender year
xtline ln_inc, overlay ///
ytitle(Mean Log-Transformed Annual Income) ///
tttitle(Wave (1 = 2006))

* Fit multilevel model of interest.
use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrslong.dta", clear

gen yrs06sq = yrssince06*yrssince06
egen unique_secu = group(stratum secu)

mixed ln_inc c.yrssince06##i.gender c.yrs06sq##i.gender ///
  i.stratum [pweight = levellwgt_r] ///
  || newid_num: yrssince06 yrs06sq, ///
  variance_cov(unstruct) ///
  pweight(level2wgt) pwscale(size) ///
  vce(cluster unique_secu)

* Error Covariance Modeling
use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrswide.dta", clear

gen ln_inc1 = ln(H8ITOT + 1)
gen ln_inc2 = ln(H9ITOT + 1)
gen ln_inc3 = ln(H10ITOT + 1)
gen ln_inc4 = ln(H11ITOT + 1)

gen comp = (ln_inc1 != . & ln_inc2 != . & ln_inc3 != . ///
& ln_inc4 != .)

```

```

* Fit a response propensity model predicting the
* probability of providing income data at all 4 waves.

* Modal imputation of missing covariate values.
replace selfrhealth_06 = 3 if selfrhealth_06 == .
replace marcat_06 = 2 if marcat_06 == .
replace diabetes_06 = 0 if diabetes_06 == .
replace arthritis_06 = 1 if arthritis_06 == .
replace racecat = 2 if racecat == .
replace edcat = 2 if edcat == .

svyset secu [pweight = kwgtr], strata(stratum)
svy: logit comp ln_incl i.selfrhealth_06 age_06 ///
      i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat, p

* Adjust baseline sampling weights in 2006 by inverse of
* predicted probability of complete data.
gen indwgt = kwgtr * (1 / phat)

* Set cluster-specific weight equal to 1 (HRS SECU weights * not provided in the public-use data).
egen newclust = group(stratum secu)
gen clustwgt = 1

* Create unique person ID.
egen newid = concat(hhid pn)
destring newid, gen(newid_num)

* Only keep cases with complete data.
keep if comp == 1

* Only keep variables of interest.
keep newid_num newclust gender ln_incl-ln_inc4 indwgt ///
clustwgt

* Reshape data to vertical format.
reshape long ln_inc, i(newid_num) j(wave)

* Create unique indicators for each wave.
gen wave1 = (wave == 1)
gen wave2 = (wave == 2)
gen wave3 = (wave == 3)
gen wave4 = (wave == 4)

* Compute constant variable for Veiga et al. macro.
gen cons = 1

* Recode gender.
gen male = (gender == 1)

* Run program pwigls_genlin_adcv_modAV1.do
cd "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code"

do "pwigls_genlin_adcv_modAV1.do"

sort newclust newid_num wave

gl wav wave1 wave2 wave3 wave4
gl x wave1 wave2 wave3 wave4 male

keep if newid_num < 20000000

mata: pwigls_genlin_adcvw4toep1("$x", "ln_inc", "$wav", "cons", "newclust", "clustwgt", "indwgt")
* end of error covariance model

*****
* Weighted GEE Modeling

use "P:\ASDA3\Data Sets for Analysis Examples and Stata R Code\hrswide.dta", clear

```

```

* Modal imputation of missing covariate values.
replace selfrhealth_06 = 3 if selfrhealth_06 == .
replace marcat_06 = 2 if marcat_06 == .
replace diabetes_06 = 0 if diabetes_06 == .
replace arthritis_06 = 1 if arthritis_06 == .
replace racecat = 2 if racecat == .
replace edcat = 2 if edcat == .

gen ln_inc06 = ln(H8ITOT + 1)
gen ln_inc08 = ln(H9ITOT + 1)
gen ln_inc10 = ln(H10ITOT + 1)
gen ln_inc12 = ln(H11ITOT + 1)

* Compute response indicator for 2008.
gen resp08 = 1 if ln_inc08 != .
replace resp08 = 0 if ln_inc08 == .
tab resp08, miss

* Response propensity model (2008).
svyset secu [pweight = kwgtr], strata(stratum)
svy: logit resp08 ln_inc06 i.selfrhealth_06 age_06 ///
      i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat08, p

* Compute response indicator for 2010.
gen resp10 = 1 if ln_inc10 != .
replace resp10 = 0 if ln_inc10 == .
tab resp10, miss

* Response propensity model (2010), respondents in 2008.
svyset secu [pweight = kwgtr], strata(stratum)
svy, subpop(if resp08 == 1): logit resp10 ln_inc06 ///
      i.selfrhealth_06 age_06 ///
      i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat10_11, p

* Response propensity model (2010), nonrespondents in 2008.
svyset secu [pweight = kwgtr], strata(stratum)
svy, subpop(if resp08 == 0): logit resp10 ln_inc06 ///
      i.selfrhealth_06 age_06 ///
      i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat10_10, p

* Compute response indicator for 2012.
gen resp12 = 1 if ln_inc12 != .
replace resp12 = 0 if ln_inc12 == .
tab resp12, miss

* Response propensity model (2012), 111 pattern.
svyset secu [pweight = kwgtr], strata(stratum)
svy, subpop(if resp08 == 1 & resp10 == 1): logit resp12 ///
      ln_inc06 i.selfrhealth_06 age_06 ///
      i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat12_111, p

* Response propensity model (2012), 110 pattern.
svyset secu [pweight = kwgtr], strata(stratum)
svy, subpop(if resp08 == 1 & resp10 == 0): logit resp12 ///
      ln_inc06 i.selfrhealth_06 age_06 ///
      i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat12_110, p

* Response propensity model (2012), 101 pattern.
svyset secu [pweight = kwgtr], strata(stratum)
svy, subpop(if resp08 == 0 & resp10 == 1): logit resp12 ///
      ln_inc06 i.selfrhealth_06 age_06 ///
      i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat12_101, p

```

```

* Response propensity model (2012), 100 pattern.
svyset secu [pweight = kwgtr], strata(stratum)
svy, subpop(if resp08 == 0 & resp10 == 0): logit resp12 ///
    ln_inc06 i.selfrhealth_06 age_06 ///
    i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat12_100, p

* Compute cumulative response probabilities.

gen cumprob1 = 1

gen cumprob2 = phat08 if resp08 == 1
replace cumprob2 = 1-phat08 if resp08 == 0

gen cumprob3 = phat08 * phat10_11 ///
    if resp08 == 1 & resp10 == 1
replace cumprob3 = phat08 * (1-phat10_11) ///
    if resp08 == 1 & resp10 == 0
replace cumprob3 = (1-phat08) * phat10_10 ///
    if resp08 == 0 & resp10 == 1
replace cumprob3 = (1-phat08) * (1-phat10_10) ///
    if resp08 == 0 & resp10 == 0

gen cumprob4 = phat08 * phat10_11 * phat12_111 ///
    if resp08 == 1 & resp10 == 1 & resp12 == 1
replace cumprob4 = phat08 * phat10_11 * (1-phat12_111) ///
    if resp08 == 1 & resp10 == 1 & resp12 == 0
replace cumprob4 = phat08 * (1-phat10_11) * phat12_110 ///
    if resp08 == 1 & resp10 == 0 & resp12 == 1
replace cumprob4 = phat08 * (1-phat10_11) * (1-phat12_110) ///
    if resp08 == 1 & resp10 == 0 & resp12 == 0
replace cumprob4 = (1-phat08) * phat10_10 * phat12_101 ///
    if resp08 == 0 & resp10 == 1 & resp12 == 1
replace cumprob4 = (1-phat08) * phat10_10 * (1-phat12_101) ///
    if resp08 == 0 & resp10 == 1 & resp12 == 0
replace cumprob4 = (1-phat08) * (1-phat10_10) * phat12_100 ///
    if resp08 == 0 & resp10 == 0 & resp12 == 1
replace cumprob4 = (1-phat08) * (1-phat10_10) * (1-phat12_100) ///
    if resp08 == 0 & resp10 == 0 & resp12 == 0

* Compute overall probability of response pattern.
gen cumprob_case = cumprob4

* Reshape data set, keeping variables of interest.
gen ln_inc1 = ln(H8ITOT + 1)
gen ln_inc2 = ln(H9ITOT + 1)
gen ln_inc3 = ln(H10ITOT + 1)
gen ln_inc4 = ln(H11ITOT + 1)

* Baseline sampling weight for individual from 2006.
gen basewgt = kwgtr

keep hhid pn gender marcat_06 diabetes_06 arthritis_06 ///
    racecat edcat ln_inc1-ln_inc4 secu stratum basewgt ///
    cumprob_case

reshape long ln_inc wgt, i(hhid pn) j(year)

* Form a unique ID variable.
egen newid = concat(hhid pn)
destring newid, gen(newid_num)

* Compute the case-specific weights for weighted GEE.
gen casewt = basewgt * (1 / cumprob_case)

* Compute measure of years since 2006, and squared version.
gen yrssince06 = 0 if year == 1
replace yrssince06 = 2 if year == 2
replace yrssince06 = 4 if year == 3
replace yrssince06 = 6 if year == 4

```

```
gen yrs06sq = yrssince06*yrssince06
```

```
* Fit weighted GEE model.
```

```
xtset newid_num year
```

```
xtgee ln_inc c.yrssince06##i.gender c.yrs06sq##i.gender ///
```

```
    i.stratum [pweight = casewt], corr(uns)
```

```
estat wcorrelation
```