* Use the following to make a do-file to check/clean hsb2_miss.dta.

* check_hsb2.do

version 9.2
capture log close
set more off
log using hsb2_check, replace
use hsb2_miss, clear

* Part I
* Point of departure: Do any of the summary values (#obs, mean, sd, min,
* max) look questionable in terms of what you expect to find? See Allison,
* _Multiple Regression: A Primer_, pages 28-29.

summarize

* Are there duplicate observations?
* Check id.
sort id
duplicates report id
duplicates list id
duplicates drop in ...

* Are there missing values? See Allison, page 29. nmissing and mvpattern are
* downloaded commands.
count
nmissing
mvpat id-socst, skip
egen mvals=rmiss(_all)
tab mvals
list id female ses race if mvals==...
* Is there an id &/or socioeconomic profile for each level of missing values, & for
* missing values in general? Explore other variables, too, including test scores.

gen mv1=(mvals==1)
tab mv1
gen mv2=(mvals==2)
tab mv2
* Do for each level of missing values.
tab mv1 mv2, chi2
tab mv1 mv..., chi2
* Do any statistically significant patterns of missing values emerge?

* Are there coding errors?
* Check variable labels:
labelbook, problem

* Check numeric coding of categorical variables.
assert id>=1 & id<=200
assert female==0 | female==1 if female<.
codebook female
assert race>=1 & race<=4 if race<.
codebook race
assert ses>=1 & ses<=3 if ses<.
codebook ses
assert schtyp==0 | schtyp==1 if schtyp<.
codebook schtyp
assert prog>=1 & prog<=3 if prog<.
codebook prog

* For quantitative variables with less than about 20 values, check that their
* distributions make sense. E.g., for years of schooling in the U.S., check that the
* variable’s frequencies are greatest at 12(high school), 14(associate’s degree), and
* 16(college degree): tab educ [this variable is not in the data set].

- For quantitative variables with more than about 20 values, check that there
  are no * extremely unusual or impossible values, such as negative
  achievement test scores. * Let’s say that on these achievement scores it’s
  impossible to score less than 20 or * more than 80.

Assert read>=20 & read<=80 if read<.
Codebook read, mv
assert write>=20 & write<=80 if write<.
Codebook write, mv
assert math>=20 & math<=80 if math<.
Codebook math, mv
assert science>=20 & science<=80 if science<.
Codebook science, mv
assert socst>=20 & socst<=80 if science<.
Codebook socst, mv

* Part II
* Quantitative variables
* univariate: do for each quantitative variable.

kdensity read, norm
gr box read
su read, d

* bivariate.

gr matrix read-socst, half
corr read-socst

* Categorical variables
* univariate: do for each categorical variable.

tab1 female-prog, miss plot

* bivariate: cross-check each categorical variable with other key categorical
* variables, including create cross-check variables where appropriate (e.g., hhead
* with adult, married with adult; emp with labor force participation, emp with
* unemp):
* The following are not in this data set but serve as examples. Create a variable that=1 if there’s a probable error:
* Confirm that all married observations are adults:
* gen errormarr=1 if marr==1 & adult==0
* tab errormarr
* list id age if errormarr==1
* tab marr if marr==1 & adult==0
* list id marr age if marr==1 & adult==0
* Confirm that, according to a questionnaire’s instructions, only hhead (household heads) answered the question about hinc (household income). In this example, we check if a honhead has a non-missing answer for hinc:
* gen errorhinc=1 if hhead==0 & hinc<.
* tab errorhinc
* list id hhead hinc if errorhinc==1

* Quantitative & categorical variables
* bivariate: cross-check each quantitative variable with key categorical variables.

tabstat read -socst, by(female) stats(mean med sd iqr min max)

* Do the following for each quantitative variable with key categorical variables:

sort female
gr box read, over(female, total)
ttest read, by(female)
oneway read ses, tab bonf

* multivariate: do a scatterplot for each quantitative variable with a key categorical variable.

scatter read write || qfit read write, by(female, total)

* Do logistic or OLS regressions using key independent variables:

xi: logistic schtype female white i.ses, nolog
reg math schtype female white i.ses

log close
exit