

* Stata OLS Regression Example.doc: Here's an example of regression analysis in Stata. The example uses the UCLA-ATS dataset hsb2.dta. If any of the commands don't work, download them via Stata's 'findit' command or via 'Help/STB & User-written Programs' (which you select by clicking 'help' on the top tool bar). April 2008.

* See 'Explanatory Variables in OLS Regression.doc'.

*** Open, describe & summarize data set, & save the listwise observations as a new data set**

```
use hsb2, clear
```

```
d
su, d
```

* Create the dummy variable 'complete', which contains only observations with non-missing data (i.e. listwise or pairwise data, which is what regression analysis uses) (even though in this particular data set there are no non-missing data).

```
mark complete
markout complete science female race ses schtyp prog read write math socst
tab complete
```

```
keep if complete==1
save complete_dataset
```

```
d
su
```

* Save the 'complete data' as a new data set, thus avoiding having to type 'if complete==1' repeatedly.

*** Note: Do the following only after thoroughly checking & cleaning the data set, including systematic univariate, bivariate, & multivariate exploratory analysis. This should include the following (or other) checks for curvilinearity in regard to each explanatory variable:**

```
qfitci scienc read
lowess science read [help lowess]
scatter science read || lowess science read, lcolor(red) || lfit science read, lcolor(blue)
scatter science read, by(female) || lowess science read, lcolor(red) || lfit science read,
lcolor(blue)
```

* mrunning to see lowess graph of dv with each iv, holding constant the other iv's. The decision could be to categorize a quantitative iv.

```
xi:mrunning science read write math socst female i.race i.ses [download 'mrunning']
```

```
locpoly science read
```

sparl science read [download 'sparl']
sparl science read, logy
sparl science read, logx
sparl science read, logy logx
sparl science read, quad

* 'boxcox' to examine whether dv needs to be transformed: theta = 1.0 don't transform dependent variable; +.5, square root of dv, 0=natural log transform of dv, -.5=reciprocal square root of dv, -1.0=reciprocal transform of dv (compare results to ladder dv, but don't do any of these unless they make sense substantively)
boxcox reg science read write math socst

* 'boxtid' to explore possible transformations of explanatory variables. Examine nlinear dev p= . The decision could be to categorize a quantitative explanatory variable.
boxtid science read write math socst female race2 race3 race4 ses2 ses4

* 'Fractional polynomials' & fracplot to evaluate whether a polynomial transformation will improve model. If a transformation is suggested, do a lowess plot. The decision could be to categorize a quantitative explanatory variable.
fracpoly regress science read write math socst female race2 race3 race4 ses2 ses4, compare
fracplot read
fracplot write
fracplot math
fracplot socst

* Format a series of final regression models in standard publication format (for options type 'help estimates'). See options for 'estimates table'. Use a word processor to edit tables for publication.

```
reg science read write  
eststo  
reg science read write math socst  
eststo  
xi3:reg science read write math socst female i.race i.ses  
eststo  
esttab, se
```

- **Some crucial issues: don't estimate regression models until you've examined & documented the bivariate (& to the extent possible, multivariate) relationships among all the variables & for the y/x relationships—overall pattern & striking deviations (i.e. form, direction & strength, including outliers). You may have to transform or do other manipulation of the variables before (or after) estimating the regression models.**
 - **Keep in mind that multiple regression analysis describes the *joint* relationships of explanatory variables to an outcome variable.**
 - ***Therefore bivariate linearity or nonlinearity & bivariate correlations do not guarantee that the relationships will be linear or that explanatory variables will test significant or not in a multiple regression model.***
-

*** Example: building & assessing a regression model**

*** Testing for model & variable significance:**

```
xi:reg science female i.race i.ses schtyp i.prog read write math socst
eststo m1
esttab, se
```

* Note: xi:reg i.catvar is a convenient way to get Stata to treat a multilevel categorical variable as nominal rather than ordinal. See also 'findit postgr' which, by using xi3 instead of xi, enables use of 'postgr' to graph interactions of categorical with quantitative vars.

```
test _lrace_2 _lrace_3 _lrace_4
test _lses_2 _lses_3
test _lprog_2 _lprog_3
```

* Test for the joint significance of the levels of each multilevel categorical variable, in order to reduce the Type I-error risks of conducting many t-tests. But how we treat the results depends on the substantive/theoretical purpose of our research.

* Note that race's categories test jointly significant & prog's categories test jointly significant, even though their individual categorical-level t-tests are not always significant (& in fact never are for prog2-3). So, drop ses, whose categories test jointly insignificant (& drop schtyp as well, since it tests insignificant), but check to see if the model would be stronger & more parsimonious with 'white' instead of race (because the white category tests significant) & 'academic' in place of program (because the academic category tests significant)

* Here's model 2, based on the above results:

```
xi: reg science female i.race i.prog read write math
eststo m2
```

* Here's model 3:

```
gen white=(race==4)
tab white
la var "White"
gen academic=(prog==2)
tab prog
la var academic "Academic program"
```

```
reg science female white academic read write math
eststo m3
test white academic
```

* White & academic test jointly significant, but so did i.race & i.prog. Note that adjusted r2 doesn't change notably, but let's stick with model 3 because of its parsimony.

```
esttab m1 m2 m3, se
```

* Next we'll submit model 3 to the various diagnostic tests.

*** Model specification test**

* 'linktest': tests for functional fit (help linktest). 'ovtest': tests for omitted variables (help ovtest). See also Stata Manual and do Google search for more information on linktest and ovtest.

help linktest
help ovtest

reg science female white academic read write math

linktest

estat ovtest

rvfplot, line(0) [residual vs. fitted values plot] or: ovfplot

* The test results are fine. Note: If there were serious problems with linktest or (perhaps especially) ovtest, the problems typically should be resolved before moving on. Sometimes problems with linktest are resolved by adding omitted variables or transforming variables, even if ovtest doesn't suggest doing so. The Stata manual discusses ovtest as crucial to get right.

* To explore problems with linktest or ovtest, consider using rvfplot options such as 'by(female) ml(id)', 'ml(female)', 'ml(id)' & ml(yhat)' The latter (i.e. the predicted values of y) is based on the command: predict yhat if e(sample).

*** Check for multicollinearity:**

help vif

findit collin [downloadable alternative to vif]

reg [repeats last model estimated]

estat vif

* Check for $vif > 10$ or so, or tolerance $< .1$ or so (although the literature cites various standards). There's no problem here.

* Alternative: collin y x1 x2 xk (downloaded command). Among other measures, 'condition' is a measure of global instability of predictors. Check for $condition > 10$ or so.

*** Check if residuals are normally distributed:**

predict rstu, rstu

histogram rstu, norm [or: hist rstu, norm; qnorm rstu, grid; gr box rstu; gr7 rstu, hist;
gr7 rstu, box]

* Graphing the residuals indicates a tight, normal distribution.

*** Check if variance is constant (i.e. homoscedastic):**

* Residual vs. fitted plot (as previously displayed)

rvfplot, yline(0) [or: ovfplot]

* Various tests for non-constant variance

help hetttest

help szroeter

estat hetttest

estat hetttest, rhs mt(sidak)

[also helpful: rdplot, group(3)]

estat szroeter

estat szroeter, rhs mt(sidak)

* Residual vs. predictor plots; and linearity checks in multivariate perspective (see also outliers/influence below)

```
rvpplot read, yline(0)
avplot read
rvpplot write, yline(0)
avplot write
rvpplot math, yline(0)
avplot math
```

* There are no problems of non-constant variance. What steps might we take if there were?

*** Check for outliers/influence:**

* leverage-residual squared plot:

```
help lvr2plot
```

```
lvr2plot, ml(id)
```

* The results look fine. Influential values would appear in the upper right portion of the table.

* avplots (to suggest not only univariate but also *multivariate* sources of influence):

```
set textsize 170 [makes text size larger]
```

```
avplots
```

```
avplot read, ml(id)
```

```
avplot write, ml(id)
```

```
avplot math, ml(id)
```

*id#150 & 167 surface as the most outlying observations, but they seem to pose no problems.

* Here's a way to detect if an observation is an outlier in more than one explanatory variable (i.e. if it's a multivariate outlier). See 'findit hadimvo'. This procedure includes a significance test (default=.05). Assess the quantitative explanatory variables only:

```
findit hadimvo [download]
```

```
help hadimvo
```

```
hadimvo read write math, gen(outlier) p(.05)
```

```
list id read write math female white prog if outlier==1
```

```
su read write math if outlier==1
```

* Let's obtain, summarize & graph the "influence" statistics. See Stata manual on when to use 'if e(sample)' ('estimation sample,' which isn't necessary in this case):

```
predict rstu if e(sample), rstu [this was predicted earlier]
```

```
predict h if e(sample), hat
```

```
predict d if e(sample), cooks
```

```
dfbeta [for all the explanatory variables]
```

```
su rstu-Dfsocst [or: su ..., detail; or: univar ... ]
```

* While rstu has one or more outliers, everything else looks quite good. This is consistent with lvr2plot & avplots.

* For the purpose of this exercise, we'll restrict in-depth examination of the influence diagnostics to cooksd ('d'). Look for values $> |d| = 1$.

```
scatter d id, yline(-1 1) ml(id)
sort d
l id d read write math if |d|>1
```

* None of the 'd's' approximate the standard ' $> |1|$ ' level for assessing 'd's' influence, but lots of them do exceed the large-sample-adjusted standard ($4/_N$). While our battery of diagnostics indicates no reason to suspect a problem, let's re-estimate the regression model excluding these d-observations (if $d < \dots$) & compare the results to the model with all the observations included.

```
reg science female white academic read write math if d<4/_N
reg science female white academic read write math
```

* Of course there needs to be some serious problem & well-reasoned argument for excluding observations. There's no need to exclude any observations in this case.

* **Test the validity of the assumption of uniform slope coefficients:** estimate a model or series of models that interact a key explanatory variables with the other explanatory variables, or selected other explanatory variables, & conduct a nested model test. In this case, let's explore the possibility that the slope coefficients are not uniform for females vs. males.

```
xi:reg science female i.female*white i.female*academic i.female*read i.female*write
i.female*math
```

```
. testparm female - _IfemXmath_1
```

* The nested model test is insignificant. I also did interaction models using race & prog, the nested model tests of which were insignificant. Hence we fail to reject the null hypothesis of uniform slopes for gender, race-ethnicity, & program.

* **Final model:**

Consider using robust standard errors in final model. Check to see if there's a difference in the pattern of significance with & without 'robust'. You can't go wrong using 'robust':

```
reg science female white academic read write math
reg science female white academic read write math, robust
```

* **Conclusions:**

* We can conclude that this regression model turned out quite well. We could re-estimate it on randomly selected sub-samples of the data set, but the small-n for each subsample is unlikely to prove helpful. We want eventually to re-estimate the model on another data set.

* **Make a publication-style results table, using esttab or outreg2:** see the course document 'Making Working and Publication Tables in Stata'.

* **Post-estimation tools to make & graph predictions:**

Let's use some 'post-estimation' tools to make & graph predictions from the model. Let's start with Stata's most basic way of making post-estimation y-predictions, "lincom" ("linear

combination"). Recall that any predictions must be within the range of the values of a data set's variables, because the model only applies to this sample of variables & values:

```
reg science female white academic read write math, robust
```

```
lincom read*45 + write*45 + math*45 + female  
lincom read*45 + write*45 + math*45 - female  
lincom _cons read*45 + write*45 + math*45 + female  
lincom _cons read*45 + write*45 + math*45 - female
```

```
lincom read*65 + write*65 + math*65 + female  
lincom read*65 + write*65 + math*65 - female  
lincom _cons read*65 + write*65 + math*65 + female  
lincom _cons read*65 + write*65 + math*65 - female
```

* Another such command is "adjust," which requires that the model include at least one explanatory categorical variable:

```
adjust read=45 write=45 math=45, by(female)  
adjust read=65 write=65 math=65, by(female)
```

* Let's look briefly at "postgr3," a set of post-estimation graph commands (which require that the regression model begin with 'xi3' & include one or more categorical explanatory vars). postgr graphs the specified variables, holding the other variables at their means or other specified levels (view help postgr, including for interaction terms).

```
xi3: reg science female white academic read write math, robust
```

```
postgr3 math  
postgr3 math, by(female) gr(s(pd))  
postgr3 math, by(female) x(read=45 write=45) gr(s(pd))  
postgr3 math, by(female) x(read=65 write=65) gr(s(pd))
```

* We alternatively can use prtab & prvalue to make linear predictions of the science achievement scores by varying the combination of independent variables, holding other variables constant at their means or other specified levels (see Long/Freese). Be wary of x-variable outlying values, however.

```
reg science female white academic read write math, robust
```

```
prtab female  
prtab white, brief  
prtab female white, brief
```

* Let's test the relative effects of 'explanatory variable' achievement score increases on predicted science scores by female versus male students.

```
prvalue, x(female=1 read=45 write=45 math=45) save brief  
prvalue, x(female=1 read=65 write=65 math=65) dif brief
```

```
prvalue, x(female=0 read=45 write=45 math=45) save brief  
prvalue, x(female=0 read=65 write=65 math=65) dif brief
```

* Finally, we can graph, e.g., the predicted science achievement scores of females & males (obtained via prtab & prvalue) holding their math achievement scores constant at 40, 50 & 60 (see Long/Freese).

```
prgen female, x(math=45) gen(fem40) n(11)
prgen female, x(math=55) gen(fem50) n(11)
prgen female, x(math=65) gen(fem60) n(11)
```

```
la var m45xb "math 45"
la var m55xb "math 55"
la var m65xb "math 65"
```

```
gr m45xb m55xb m65xb m65x, c(III) s(opd) xla(0 1) yla(44 47 to 56) b1("Predicted Female &
Male Science Scores by Math Scores") b2(male=0 female=1) l2("Predicted Science Scores")
g(3)
```

How to graph a yhat confidence interval

```
reg science math
```

* Hard way:

```
predict yhat if e(sample)
predict se, stdp
di invttail(199, .05/2)                answer: 1.972
gen lo = yhat - (1.972*se)
gen hi = yhat + (1.972*se)
```

* Easy way

```
scatter science yhat lo hi math, c(. |||) m(O i i i) sort
```

How to graph a regression model with a squared term

```
gen expersq=exper^2                [Wooldrige: WAGE1.dta]
su exper expersq
```

```
reg lwage exper expersq
```

* Hard way

```
predict yhat2 if e(sample)
predict se, stdp
di invttail(525,.05/2)             answer: 1.972
gen lo = yhat2 - (1.97*se)
gen hi = yhat2 + (1.97*se)
scatter lwage yhat2 lo hi exper, c(. |||) m(O i i i) sort
```

* Easy way:

twoway qfitci l wage exper [automatically fits yhat with squared-exper, with confidence interval]

* Check to see at what level of exper(ience) the curvature begins & whether or not this makes substantive or theoretical sense.

How to impute missing values with a probabilistic component (see 'view help uvis' or 'view help mvis' for a description of the assumptions & command options). See the book by Paul Allison and the article by Gary King et al. concerning the relevance of imputation methods.

Note: findit mim [to use multiply imputed data]

To impute a missing y-value:

```
uvis regress y x1 x2 x3, gen(y_new) seed(123)
```

But every missing value for x1-x3 will create a missing value in y, even if the y-value wasn't originally missing.

If that happens, after running uvis try the following procedure to replace the new missing y-values with the original non-missing y-values (compliments of Stata listserve).

```
replace y_new = max(y, y_new)
```

To impute missing x-values:

```
mvis x1 x2 x3 x4 using filename, m(10) cc(x6 x8) cmd(x1 x2:logit, x3:ologit, x4:regress)  
cmd(cmdlist) genmiss(m_) id(pid) seed(123)
```