# Stata Tutorial: Session II

Regression

- Estimation Commands
- Dummy variables
- Incorporating additional commands
- Basic table of results

**Estimation Commands**

OLS Estimator: `regress`

Syntax:

```
regress depvar [indepvars] [if] [, options]
```

Important options are how standard errors are calculated:

| | |
|---|---|
| `robust` | Robust to heteroskedasticity & autocorrelation |
| `cluster(var)` | Block diagonal Variance-Covariance matrix |
| `vce(bootstrap)` | Repeated estimation with random samples |
| `vce(jackknife)` | Repeated estimations dropping one observation each time |

**Estimation Commands**

OLS Estimator: `regress`

Syntax:

`regress depva` ... `tions]`

Panel Data ID variable

(country, individual, etc.)

Important optio... ...ted:

| | |
|---|---|
| `robust` | ...asticity & autocorrelation |
| `cluster(var)` | Block diagonal Variance-Covariance matrix |
| `vce(bootstrap)` | Repeated estimation with random samples |
| `vce(jackknife)` | Repeated estimations dropping one observation each time |

**Estimation Commands**

OLS Estimator: `regress`

Syntax:

```
regress depvar [indepvars] [if] [, options]
```

Importa… …are calculated:

`r…` …skedasticity &

`clust…` …nal Variance-Covariance Matrix

`vce(bootstrap)` Repeated estimation with random samples

`vce(jackknife)` Repeated estimations dropping one observation each time

`vce` = variance-covariance estimator

## Estimation Commands

Estimation results saved by `reg`:

| | |
|---|---|
| `e(N)` | number of observations |
| `e(mss)` | model sum of squares (ESS) |
| `e(df_m)` | model degrees of freedom (k) |
| `e(rss)` | residual sum of squares (RSS) |
| `e(df_r)` | residual degrees of freedom (n-k-1) |
| `e(r2)` | *R*-squared |
| `e(r2_a)` | adjusted *R*-squared |
| `e(F)` | *F* statistic |
| `e(b)` | coefficient vector |
| `e(V)` | variance-covariance matrix |
| `e(sample)` | dummy variable for in sample |

**Estimation Commands**

Things you can do with saved estimation results:

- List: `ereturn list`
- Display: `display e(N)`
  `matrix list e(V)`
  `display _b[lnpop]`
- limit sample: `br if e(sample)`
- calculate: `display e(df_m)+e(df_r)`
- save: `scalar RSS_R=e(rss)`

**Dummy Variables**

Dummy variable = 0 (not in group) or 1 (in group)

Examples:

```
gen coldwar=(year<1992)
gen Botswana=(country== "Botswana ")
reg lnUSA_TOFG lnGDP Botswana coldwar
```

For large number of dummies, could be tedious! Suppose you want a dummy for each year instead of just `coldwar`:

```
reg lnUSA_TOFG lnGDP Botswana i.year
```

**Dummy Variables**

Dummy variable = 0 (not in group) or 1 (in group)

Examples:

```
gen cold
gen                                    ana ")
reg                                    oldwar
```

`i.` = indicator variable

[`c.` = continuous variable]

For large number                       ous! Suppose you want a dummy for each year instead of just                ldwar:

```
reg lnUSA_TOFG lnGDP Botswana i.year
```

**Dummy Variables**

`i.` only works for numbers.

How could we include a dummy for each country?

➔ First convert to an id number

```
encode country, gen(cid)
reg lnUSA_TOFG lnGDP i.cid i.year
```

**Dummy Variables**

What if we want to multiply variables within regression?

“Factor variables”

# include multiplied only

## include un-multiplied and multiplied

```
reg lnUSA_TOFG lnGDP coldwar c.lnGDP#i.coldwar
reg lnUSA_TOFG c.lnGDP##i.coldwar
```

**Dummy Variables**

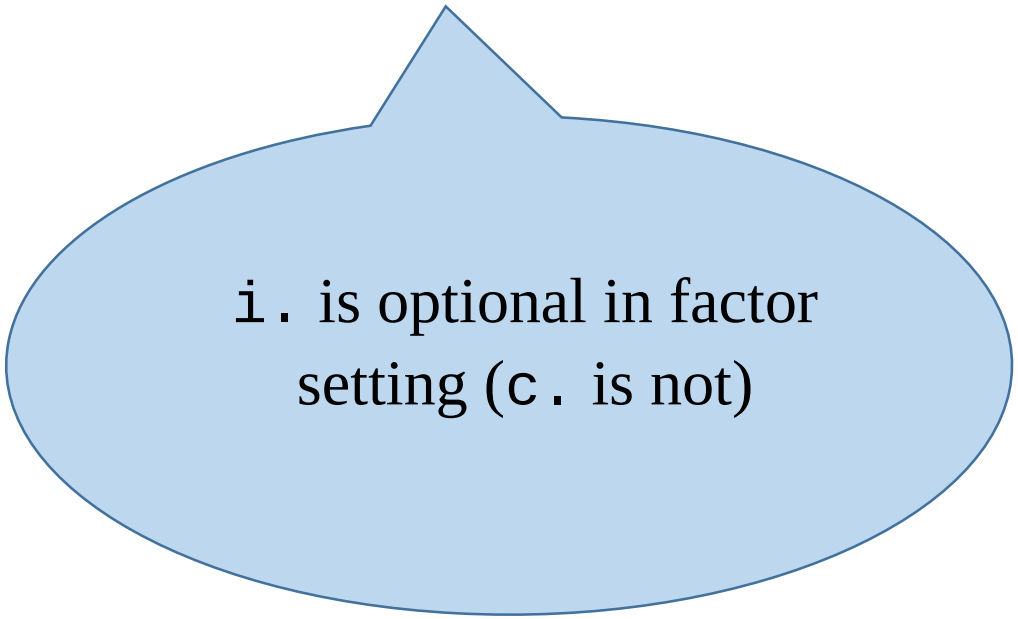
What if we want to multiply variables within regression?

"Factor variables"

# include multiplied only

## include un-multiplied and multiplied

```
reg lnUSA_TOFG lnGDP coldwar c.lnGDP#i.coldwar
reg lnUSA_TOFG c.lnGDP##i.coldwar
```



**Dummy Variables**

Why bother with factor variables?

```
gen lnGDP_cw=lnGDP*coldwar
reg lnUSA_TOFG lnGDP coldwar lnGDP_cw
```

Is this better or worse than

```
reg lnUSA_TOFG c.lnGDP##i.coldwar
```

Better:

- ➔ Stata knows relationship between all three variables
- ➔ Less clutter!

Some commands cannot handle “factor variables”—then have to use first approach.

**Time Series Operators**

If you have time series or panel data, can specify timing.

First have to tell STATA about time:

Time Series: `tsset year`

Panel: `xtset cid year`

Then use lag/forward/difference operators:

```
reg lnUSA_TOFG L.lnGDP coldwar
reg lnUSA_TOFG L2.lnGDP coldwar
reg lnUSA_TOFG L(1/3).lnGDP coldwar
reg F.lnUSA_TOFG lnGDP coldwar
reg lnUSA_TOFG D.lnGDP coldwar
reg lnUSA_TOFG LD.lnGDP coldwar
```

**Post-Estimation Commands**

Each STATA estimation command allows certain other post-estimation commands. List them by typing:

```
help regress postestimation
```

[can be run without the `estat` at the beginning]

Examples:

```
reg lnUSA_TOFG lnGDP coldwar if region==7
vif
hettest
ovtest
imtest, white
```

## Post-Estimation Commands

More Examples:

```
reg lnUSA_TOFG lnGDP coldwar if region==7
predict yhat, xb
predict ehat, resid
hist ehat, norm
#delimit ;
margins, atmeans at(lnGDP=(1(1)10))
   plot(recast(line) recastci(rarea));
```

## Post-Estimation Commands

More Examples:

```
gen lnPop=log(pop)
reg lnUSA_TOFG lnGDP lnPop coldwar
test (lnGDP=0) (lnPop=0)
test (lnGDP= -2*lnPop)

reg lnUSA_TOFG lnGDP lnPop if year==2001
avplots
avplot lnPop, mlab(cid)
```

**Basic Table of Results**

Table can:

- include results from several regression (multiple columns)
- exclude coefficients from uninteresting variables

Examples:

```
reg lnUSA_TOFG lnGDP lnPop if region==1
eststo e1
reg lnUSA_TOFG lnGDP lnPop if region==2
eststo e2
esttab e1 e2, nogaps


reg lnUSA_TOFG lnGDP lnPop i.year, cluster(cid)
eststo e4
esttab e4, nogaps keep(lnGDP lnPop)
```

## Commands Used

| | |
|---|---|
| `#delimit` | set end-of-command marker |
| `avplot` | added variable plot |
| `display` | display text or numbers |
| `encode` | create id # from text |
| `ereturn list` | list estimation results |
| `eststo` | store estimation output |
| `esttab` | create table of estimations |
| `gen` | generate new variable |
| `hettest` | heteroskedasticity test |
| `hist` | histogram |
| `imtest, white` | White heteroskedasticity test |
| `margins` | predictive margins |
| `matrix list` | display a matrix |

| | |
|---|---|
| `ovtest` | omitted variable test (RESET) |
| `predict` | predict fitted values or resid |
| `reg` | regression |
| `scalar` | generate a scalar variable |
| `test` | F-test |
| `tsset` | set time series identifier |
| `vif` | Variance Inflation Factor |
| `xtset` | set panel identifiers |
| `L.` | lag operator |
| `F.` | forward operator |
| `D.` | different operator |
| `i.` | indicator variable |
| `c.` | continuous variable |
| `#` | factor (interaction) |
| `##` | regular & factor variables |