ANOVA/Regression I
Session 1 : Simple Linear Regression

In this session we are going to work in STATA and we want to load the *sbpage* dataset. The next step is to open a *log* file to save all the text-output, which later can be loaded into a text editor or word processor. To open it click on the **Open Log** button from the toolbar and give a filename, we can call it *lab1.log* and then click on the **Open** button, a Stata Log (white) window will appear and in there all the text output will be saved. For now we can close the Stata Log window by clicking on the **X** on the top right corner of this window. You can temporarily suspend output from being written to the log by clicking on the **Close/Suspend Log** button (it is the same as the **Open Log** button), select the **Suspend log file** option and then click **OK**. Then you can open the log-file again by clicking on the **Close/Resume Log** button (it is the same as the **Open Log** button), select **Resume suspended log file** option and then click **OK**.

Saving the graph in Word:
A way to save the graph in Word is to copy the graph to the clipboard and then to import it into Word or another Windows application. Follow the below steps:
1. Display your graph in the Stata Graph window.
2. Click on the title bar of the Stata Graph window.
3. Choose **Copy Graph** from the **Edit** menu.
4. In the other Windows application, you can then choose **Paste** from the **Edit** menu.

Now we have a dataset and a log-file so let's start the analysis.

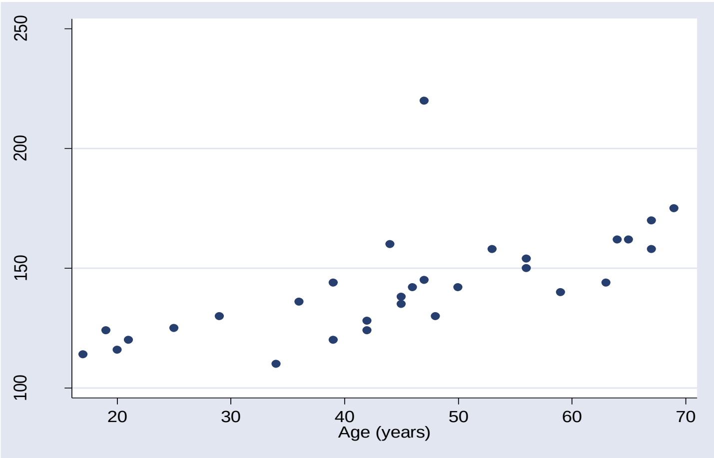
1. *Sbpage Dataset*:
To view the data you can type:
```
list
```

If the dataset is too big and you don't want to list all the observations you can view the first 10 observations with the following command:
```
list in 1/10
```

We have two variables of interest, systolic blood pressure (*sbp*) and *age*. We can set labels to the variables *sbp* and *age*:
```
label var sbp "Systolic Blood Pressure (mm Hg)"
label var age "Age (years)"
```

We want to explore the association between systolic blood pressure and age, so first we produce a *scatter diagram*:

```
scatter sbp age
```



a. From the above scatter-plot what can you tell about the relationship between *sbp* and *age* ? What would you suspect to be the sign of the slope?

To regress *sbp* on *age*, the command is the following:

```
regress sbp age


  Source |       SS       df       MS              Number of obs =      30
---------+------------------------------           F(  1,    28) =   21.33
   Model |  6394.02269     1  6394.02269           Prob > F      =  0.0001
Residual |  8393.44398    28  299.765856           R-squared     =  0.4324
---------+------------------------------           Adj R-squared =  0.4121
   Total |  14787.4667    29  509.912644           Root MSE      =  17.314

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .9708704   .2102157      4.618   0.000       .5402629    1.401478
   _cons |   98.71472   10.00047      9.871   0.000       78.22969    119.1997
------------------------------------------------------------------------------
```

b. Please find the following quantities from the output:

$\hat{\beta}_0 =$ $\qquad$ $\hat{\beta}_1 =$ $\qquad$ So the estimated least-square line is $\hat{Y} = \quad + \quad X$

$\sqrt{MSE} =$ $\qquad$ Coefficient of Determination or $R^2 =$

c. Test the null hypothesis of no linear association. What is the association between the *F-test* of the model and the *t-test* of the slope?

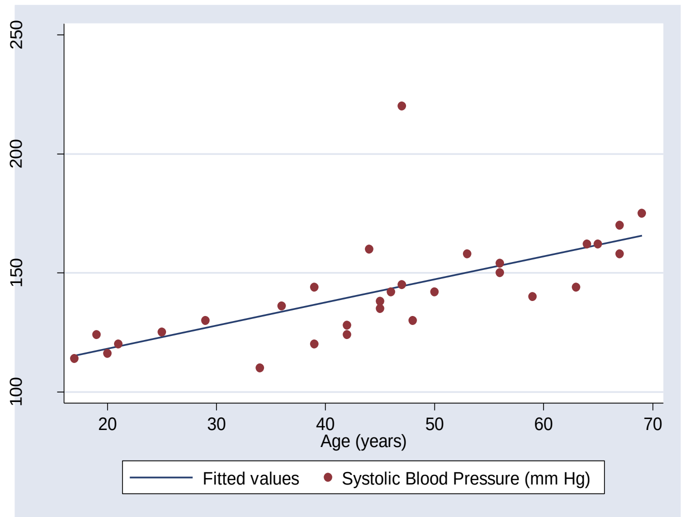
STATA by default calculates 95% confidence intervals if we would like to change them to 90% we should add the option *level(#)* after the regress command as following:

```
regress sbp age, level(90)


  Source |       SS       df       MS              Number of obs =      30
---------+------------------------------           F(  1,    28) =   21.33
   Model |  6394.02269     1  6394.02269           Prob > F      =  0.0001
Residual |  8393.44398    28  299.765856           R-squared     =  0.4324
---------+------------------------------           Adj R-squared =  0.4121
   Total |  14787.4667    29  509.912644           Root MSE      =  17.314

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [90% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .9708704   .2102157      4.618   0.000       .6132659    1.328475
   _cons |   98.71472   10.00047      9.871   0.000       81.70261    115.7268
------------------------------------------------------------------------------
```

d. What did change in the output? Give the new confidence intervals of the slope and the intercept, did they become wider or narrower and why?

We can produce a graph of the estimated regression line on the scatter diagram:

twoway (lfit sbp age) (scatter sbp age)

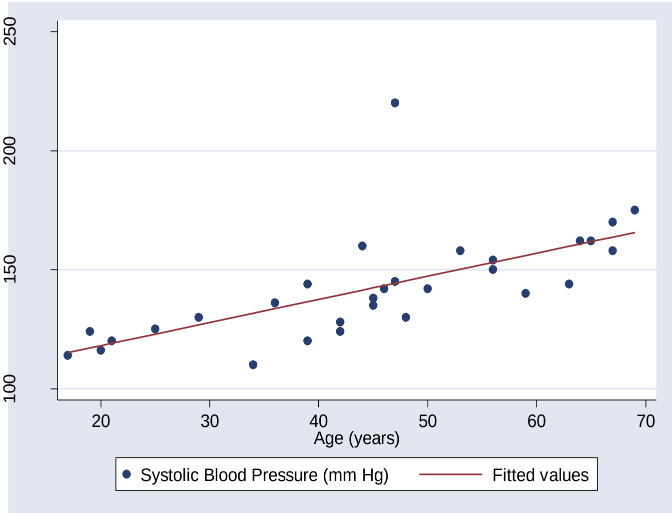


We can get the same graph using the fitted values :

predict sbphat

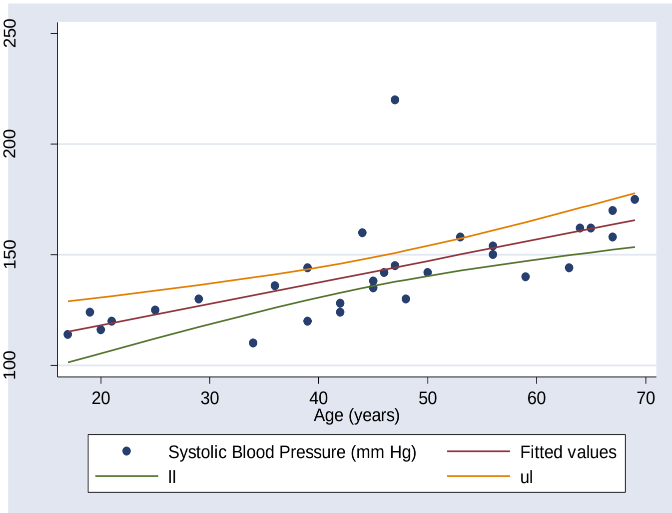
sort sbphat

```
twoway (scatter sbp age) (line sbphat age)
```



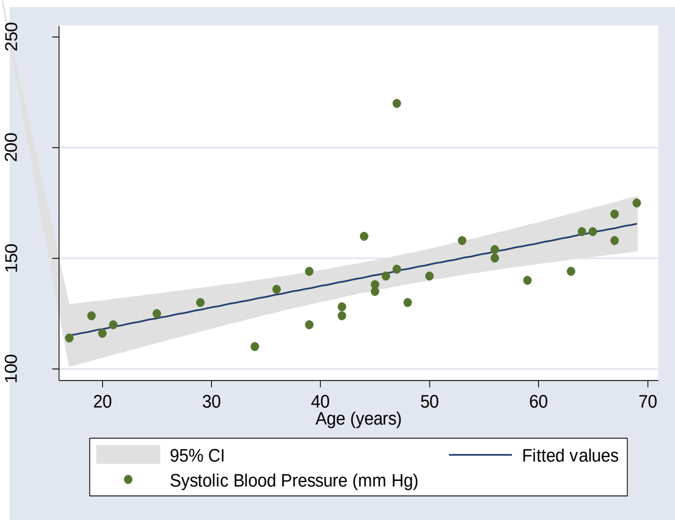
Now we want to construct 95% confidence intervals about the regression line using fitted values and their standard error ($S_{\hat{y}}$):

```
predict s, stdp
gen ul=sbphat+invttail(28,0.025)*s
gen ll=sbphat-invttail(28,0.025)*s

sort ul

twoway (scatter sbp age) (line sbphat age) (line ll age) ///
(line ul age)
```



We can also use **lfitci** command in Stata:

```
twoway (lfitci sbp age, stdp) (scatter sbp age)
```

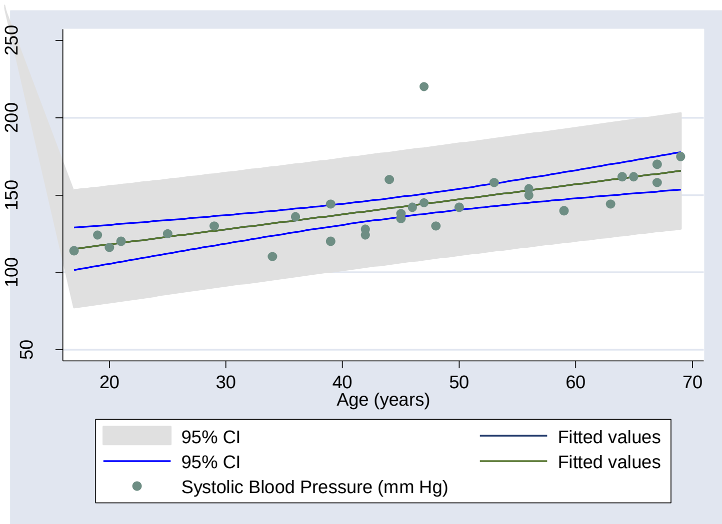


In the same way, if we want the 95% confidence intervals of a prediction you type the following set of commands using the standard error of the prediction ($S_{\hat{y}_{x_0}}$):

```
predict sf, stdf
gen ulpred=sbphat+invttail(28,0.025)*sf
gen llpred=sbphat-invttail(28,0.025)*sf
```

Next the command to get also the prediction bands in the graph follows, using **lfitci** with option **stdf** :

```
twoway (lfitci sbp age, stdf) ///
(lfitci sbp age, stdp ciplot(rline) blcolor(blue)) ///
(scatter sbp age)
```



Notice that one point, (47,220), seems quite out of place; such an observation is often called an *outlier*. One easy way to identify the outlier is to use the **"`mlabel(id)`"** option in the "**`scatter`**" command**.**

```
scatter sbp age, mlabel(id)
```

Now we want to refit the model without the outlier one way is the following :

```
regress sbp age if sbp!=220
                                            (In STATA "!=" stands for "not equal to")


  Source |       SS       df       MS              Number of obs =      29
---------+------------------------------           F(  1,    27) =   66.81
   Model |  6110.10173     1  6110.10173           Prob > F      =  0.0000
Residual |  2469.34654    27  91.4572794           R-squared     =  0.7122
---------+------------------------------           Adj R-squared =  0.7015
   Total |  8579.44828    28  306.408867           Root MSE      =  9.5633

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .9493225   .1161445      8.174   0.000       .7110137    1.187631
   _cons |   97.07708   5.527552     17.562   0.000       85.73549    108.4187
------------------------------------------------------------------------------
```

e. Please find again the following quantities from the output. What can you tell this time about the fit of the model?

$\hat{\beta}_0 =$ $\hat{\beta}_1 =$ So the estimated least-square line is $\hat{Y} = \quad + \quad X$

$\sqrt{MSE} =$ Coefficient of Determination or $R^2 =$

ANOVA/Regression I
Solutions to Session 1: Simple Linear Regression

1. *Sbpage Dataset*:

a. The relationship seems to be proportional thus the sign of the slope positive, blood pressure *increases* with age.

b. $\hat{\beta}_0 =$ 98.71. $\hat{\beta}_1 =$ 0.9709 So the estimated least-square line is $\hat{Y} = 98.71 + 0.97X$

$\sqrt{MSE}$ =17.314. The coefficient of Determination or $R^2 =$ 0.4324 or 43%

c. **critical value F:** `display invFtail(1,28, 0.05)`
p-value: `display Ftail(1,28, 21.33)`

$F=21.33 > F_{1,28,0.95}=4.20$ (or p-value=0.0001<0.05=α), thus reject the null hypothesis of no linear association between blood pressure and age. Alternatively, $T=4.618 > t_{28,0.975}=2.048$ (or p-value = 0.0001< 0.05=α) and again reject the null hypothesis.

critical value T: `display invttail(28, 0.025)`
p-value: `display 2*ttail(28, 4.618)`

The association is the following $T^2=(4.618)^2=21.33=F$.

d. Only the confidence intervals of the slope and the intercept changed.
90% CI of slope: [.6132659 1.328475]
90% CI of intercept: [81.70261 115.7268]

They became narrower since the significance level $\boldsymbol{\alpha}$ increased to 10% and thus the critical values of the t-distribution smaller($t_{28,0.95}=1.701 < t_{28,0.975}=2.048$).

```
display invttail(28, 0.05)
display invttail(28, 0.025)
```

e. $\hat{\beta}_0 =$ 97.08 $\hat{\beta}_1 =$ 0.9493 So the estimated least-square line is $\hat{Y} = 97.08 + 0.95X$

$\sqrt{MSE}$ = 9.563 Coefficient of Determination or $R^2 =$ 0.7122 or 71%

Now the fit of the model is better, since $\sqrt{MSE}$ is smaller (9.563 vs. 17.314) and R^2 is greater (43% vs. 71%), this model is more precise and more variability is explained when omitting the outlier from the analysis.