

## ANOVA/Regression I

### Session 2: Multiple Regression

In today's session we are going to work with the dataset *wgthg* where the weight (WGT) in kilograms, height (HGT) in centimetres and age (AGE) in years were recorded for a random sample of 12 nutritionally deficient children, so we are going to load the dataset *wgthgt.dta*. We are interested to describe the relationship of weight to height and age for this group of nutritionally deficient children. The first thing we are going to do is to open a log file in order to save the results, we can call it *lab2.log*. Then we are going to label the variables with the appropriate labels:

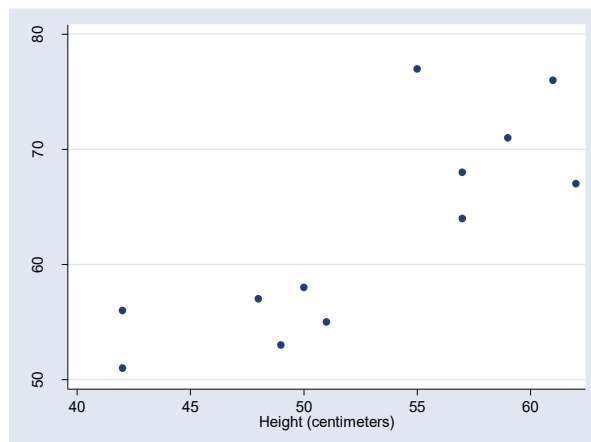
```
label var wgt "Weight (kgs)"
label var hgt "Height (centimeters)"
label var age "Age (years)"
gen age2=age*age
label var age2 "Square of Age"
```

Now we want to view the data to see how they look like:

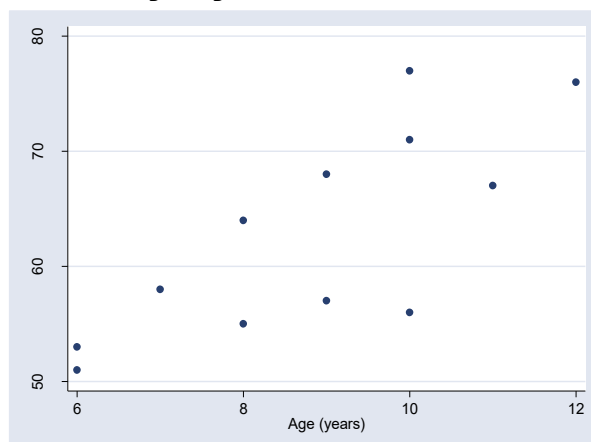
```
list
sum
```

Next we are going to produce the scatter plots to get an idea what is the relationship between these variables:

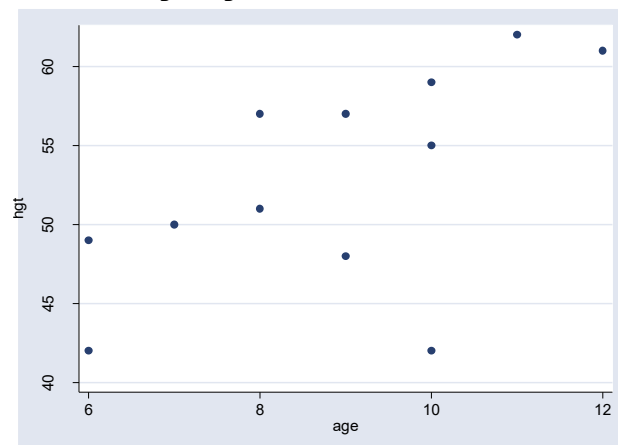
```
scatter wgt hgt
```



```
scatter wgt age
```

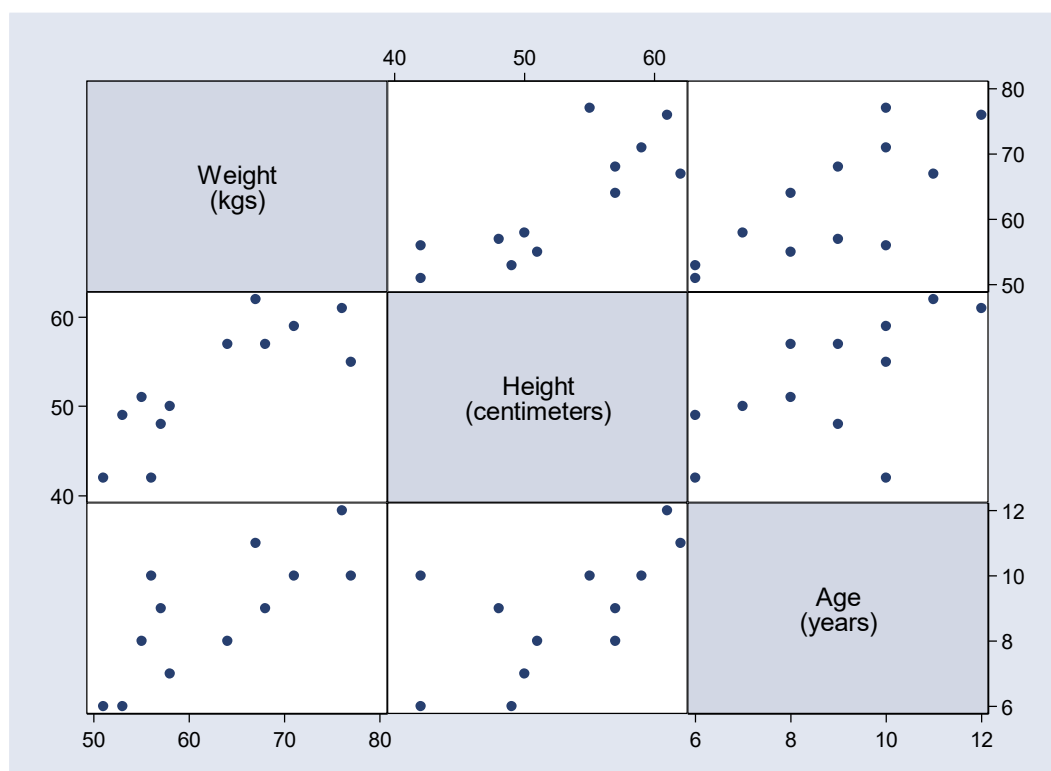


```
scatter hgt age
```



Another way to get the scatter plots of all three variables at once is to type the following command:

```
graph matrix wgt hgt age
```



The relationship of the variables seems to be positive. We can also get the estimates of the correlations by typing:

**pwcorr wgt hgt age, sig**

	wgt	hgt	age
wgt	1.0000		
hgt	0.8143	1.0000	
age	0.7698	0.6138	1.0000
	0.0034	0.0337	

So height is highly positively correlated with weight ( $r=0.81$ ), as is age ( $r=0.77$ ). Now we are going to start our modeling procedure, first we will fit the univariate models and then we are going to fit the more complex models with multiple variables in it. So we have:

### Model 1: $WGT=b_0+ b_1HGT + e$

**regress wgt hgt**

Source	SS	df	MS	Number of obs = 12		
Model	588.922523	1	588.922523	F( 1, 10)	=	19.67
Residual	299.327477	10	29.9327477	Prob > F	=	0.0013
Total	888.25	11	80.75	R-squared	=	0.6630
				Adj R-squared	=	0.6293
				Root MSE	=	5.4711

wgt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hgt	1.07223	.241731	4.44	0.001	.5336202	1.610841
_cons	6.189849	12.84875	0.48	0.640	-22.43894	34.81864

### Model 2: $WGT=b_0+ b_2AGE + e$

**regress wgt age**

Source	SS	df	MS	Number of obs = 12		
Model	526.392857	1	526.392857	F( 1, 10)	=	14.55
Residual	361.857143	10	36.1857143	Prob > F	=	0.0034
Total	888.25	11	80.75	R-squared	=	0.5926
				Adj R-squared	=	0.5519
				Root MSE	=	6.0155

wgt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	3.642857	.9551151	3.81	0.003	1.514728	5.770986
_cons	30.57143	8.613705	3.55	0.005	11.3789	49.76396

### Model 3: $WGT = b_0 + b_3(AGE)^2 + e$

**regress wgt age2**

Source	SS	df	MS	Number of obs = 12		
Model	521.932047	1	521.932047	F( 1, 10)	=	14.25
Residual	366.317953	10	36.6317953	Prob > F	=	0.0036
				R-squared	=	0.5876
				Adj R-squared	=	0.5464
				Root MSE	=	6.0524
wgt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age2	.2059716	.0545669	3.77	0.004	.0843889	.3275543
_cons	45.99764	4.76964	9.64	0.000	35.37022	56.62506

In order to get also the partial *Type I F-tests* in the following multiple models we are going to use the `anova` command:

### Model 4: $WGT = b_0 + b_1HGT + b_2AGE + e$

For previous versions of Stata (<11) the command syntax is  
**anova wgt hgt age, continuous(hgt age) sequential**  
**anova, regress**

In Stata 11 we type:

**anova wgt c.hgt c.age, sequential**

		Number of obs = 12		R-squared = 0.7800	
		Root MSE = 4.65984		Adj R-squared = 0.7311	
Source	Seq. SS	df	MS	F	Prob > F
Model	692.822607	2	346.411303	15.95	0.0011
hgt	588.922523	1	588.922523	27.12	0.0006
age	103.900083	1	103.900083	4.78	0.0565
Residual	195.427393	9	21.7141548		
Total	888.25	11	80.75		

**regress**

Source	SS	df	MS	Number of obs = 12		
Model	692.822607	2	346.411303	F( 2, 9)	=	15.95
Residual	195.427393	9	21.7141548	Prob > F	=	0.0011
				R-squared	=	0.7800
				Adj R-squared	=	0.7311
				Root MSE	=	4.6598
wgt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hgt	.722038	.2608051	2.77	0.022	.1320559	1.31202
age	2.050126	.9372256	2.19	0.056	-.0700253	4.170278
_cons	6.553048	10.94483	0.60	0.564	-18.20587	31.31197

**Model 5:  $WGT = b_0 + b_1HGT + b_3(AGE)^2 + e$**

**anova wgt c.hgt c.age2, sequential**

Number of obs = 12					
Root MSE = 4.69752					
R-squared = 0.7764					
Adj R-squared = 0.7267					
Source	Seq. SS	df	MS	F	Prob > F
Model	689.649951	2	344.824976	15.63	0.0012
hgt	588.922523	1	588.922523	26.69	0.0006
age2	100.727428	1	100.727428	4.56	0.0614
Residual	198.600049	9	22.0666721		
Total	888.25	11	80.75		

**regress**

Source	SS	df	MS	Number of obs = 12		
Model	689.649951	2	344.824976	F( 2, 9) = 15.63		
Residual	198.600049	9	22.0666721	Prob > F = 0.0012		
Total	888.25	11	80.75	R-squared = 0.7764		
				Adj R-squared = 0.7267		
				Root MSE = 4.6975		
wgt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hgt	.7259765	.2633306	2.76	0.022	.1302814	1.321672
age2	.1148016	.0537332	2.14	0.061	-.0067513	.2363546
_cons	15.11754	11.7969	1.28	0.232	-11.5689	41.80398

**Model 6:  $WGT = b_0 + b_1HGT + b_2AGE + b_3(AGE)^2 + e$**

**anova wgt c.hgt c.age c.age2, sequential**

Number of obs = 12					
Root MSE = 4.9395					
R-squared = 0.7803					
Adj R-squared = 0.6978					
Source	Seq. SS	df	MS	F	Prob > F
Model	693.060463	3	231.020154	9.47	0.0052
hgt	588.922523	1	588.922523	24.14	0.0012
age	103.900083	1	103.900083	4.26	0.0730
age2	.237856856	1	.237856856	0.01	0.9238
Residual	195.189537	8	24.3986921		
Total	888.25	11	80.75		

## regress

Source	SS	df	MS	Number of obs = 12		
Model	693.060463	3	231.020154	F( 3, 8) = 9.47		
Residual	195.189537	8	24.3986921	Prob > F = 0.0052		
Total	888.25	11	80.75	R-squared = 0.7803		
				Adj R-squared = 0.6978		
				Root MSE = 4.9395		

wt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hgt	.7236902	.2769632	2.61	0.031	.085012	1.362368
age	2.776875	7.427279	0.37	0.718	-14.35046	19.90421
age2	-.0417067	.4224071	-0.10	0.924	-1.015779	.9323659
_cons	3.438426	33.61082	0.10	0.921	-74.06826	80.94512

1. In order to get a summary of all the models please fill in the below table with the appropriate quantities:

Model	Variables Used	SSR	d.f.	SSE	d.f.	Overall F	p	R <sup>2</sup>
1	HGT							
2	AGE							
3	(AGE) <sup>2</sup>							
4	HGT, AGE							
5	HGT, (AGE) <sup>2</sup>							
6	HGT, AGE, (AGE) <sup>2</sup>							

We see that all models have a significant overall F-test, now we want to find out which model is best. For example whether adding age to the model after controlling for height will contribute significantly. For this kind of test we can use a *partial Type I F-test*. This procedure is called *variables-added-in-order* method.

2. So from model 1 and 4 find the following quantities and compute the Type I F-test and look in the model's 4 output to see where STATA gives you this result. Also compare it to the t-test of age in model 4.

$$F(AGE | HGT) = \frac{SSR(HGT, AGE) - SSR(HGT)}{SSE(HGT, AGE)/(n - k - 1)} = \frac{SSE(HGT) - SSE(HGT, AGE)}{SSE(HGT, AGE)/(n - k - 1)}$$

3. Now what about comparing model 4 and 6, whether the addition of (AGE)<sup>2</sup> to the model after controlling for both height and age contributes significantly. Find from the output of model 6 the appropriate *partial Type I F-test*,  $F((AGE)^2 | HGT, AGE)$  and the corresponding p-value and what is your conclusion.

Now we can address the same questions as above but using another procedure the *variables-added-last* regression models to get the *Type III partial F-tests*. These can be derived by running several regressions each time entering the variable in question last. To do this in STATA we still need the anova sequential command and we are going to use in each case model 6 defined previously but we are going to change the sequence of the variable added last. So we already have the multiple model (**model 6**) where we added last (AGE)<sup>2</sup> and now we need to define another two models where AGE (**model 7**) and HEIGHT (**model 8**) will be entered last.

**Model 7:  $WGT = b_0 + b_1HGT + b_3(AGE)^2 + b_2AGE + e$**

**anova wgt c.hgt c.age2 c.age, sequential**

		Number of obs = 12		R-squared = 0.7803	
		Root MSE = 4.9395		Adj R-squared = 0.6978	
Source	Seq. SS	df	MS	F	Prob > F
Model	693.060463	3	231.020154	9.47	0.0052
hgt	588.922523	1	588.922523	24.14	0.0012
age2	100.727428	1	100.727428	4.13	0.0766
age	3.41051231	1	3.41051231	0.14	0.7182
Residual	195.189537	8	24.3986921		
Total	888.25	11	80.75		

**Model 8:  $WGT = b_0 + b_3(AGE)^2 + b_2AGE + b_1HGT + e$**

**anova wgt c.age2 c.age c.hgt, sequential**

		Number of obs = 12		R-squared = 0.7803	
		Root MSE = 4.9395		Adj R-squared = 0.6978	
Source	Seq. SS	df	MS	F	Prob > F
Model	693.060463	3	231.020154	9.47	0.0052
age2	521.932047	1	521.932047	21.39	0.0017
age	4.5464612	1	4.5464612	0.19	0.6774
hgt	166.581955	1	166.581955	6.83	0.0310
Residual	195.189537	8	24.3986921		
Total	888.25	11	80.75		

Note that we didn't put the `regress` option after `anova` as before, since the regression ANOVA table for the above models is exactly the same as for model 6.

4. Find out from the above outputs including model's 6 the following sum of squares:

$$SS((AGE)^2 | HGT, AGE) =$$

$$SS(AGE | HGT, (AGE)^2) =$$

$$SS(HGT | AGE, (AGE)^2) =$$

5. The *Type III F-tests* are derived by dividing the above sum of squares by the full model mean square error  $MSE(HGT, AGE, (AGE)^2)$ . With this information compute the following *Type III F-tests* and find out where STATA gives you these tests so also report the corresponding p-values:

$$F((AGE)^2 | HGT, AGE) =$$

$$F(AGE | HGT, (AGE)^2) =$$

$$F(HGT | AGE, (AGE)^2) =$$

6. According the above results which variable would you suspect to be the most important predictor of *weight*.

Now an easier way to get immediately the partial *Type III F-tests* for all three variables in STATA is by specifying the option `partial` or by not specifying an option at all since `partial` is the default.

**anova wgt c.hgt c.age c.age2, partial**

Number of obs = 12      R-squared = 0.7803 Root MSE = 4.9395      Adj R-squared = 0.6978					
Source	Partial SS	df	MS	F	Prob > F
Model	693.060463	3	231.020154	9.47	0.0052
hgt	166.581955	1	166.581955	6.83	0.0310
age	3.41051231	1	3.41051231	0.14	0.7182
age2	.237856856	1	.237856856	0.01	0.9238
Residual	195.189537	8	24.3986921		
Total	888.25	11	80.75		

An other way is by using command `test` after `regress`:

**quietly regress wgt hgt age age2**

**test hgt**

( 1) hgt = 0
F( 1, 8) = 6.83
Prob > F = 0.0310

**test age**

( 1) age = 0
F( 1, 8) = 0.14
Prob > F = 0.7182

**test age2**

( 1) age2 = 0
F( 1, 8) = 0.01
Prob > F = 0.9238