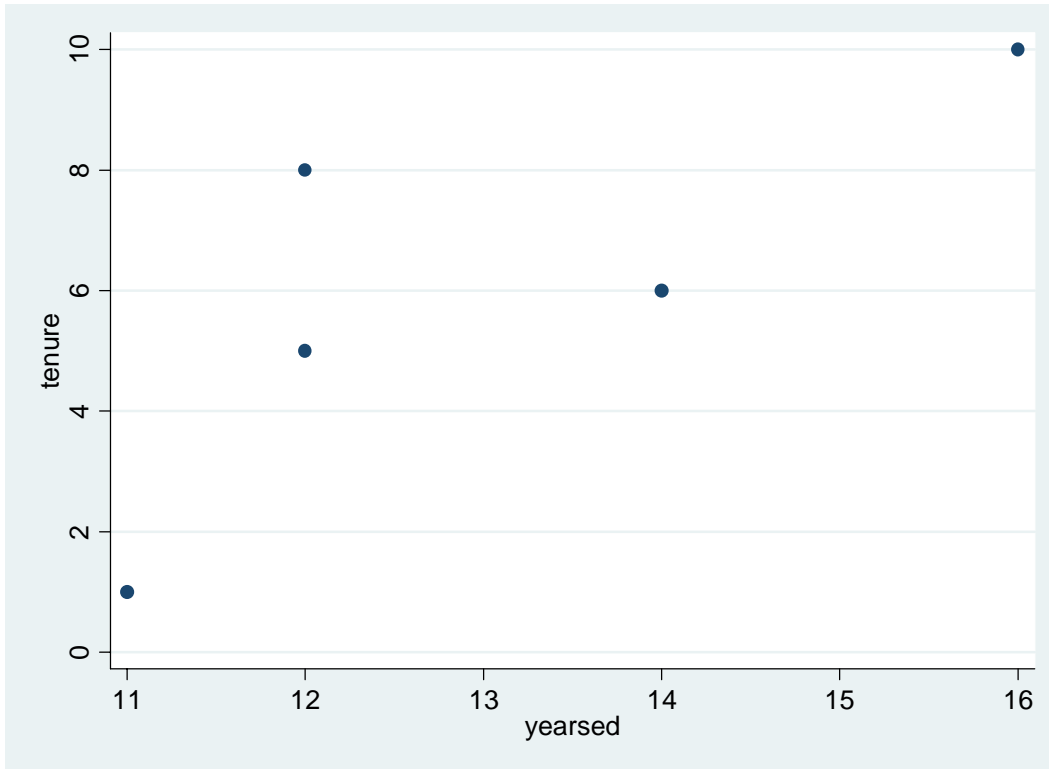


1. Stata command to obtain scatter diagram plotting the pairs of observations of Y (job tenure) and X (years of education)

(remember Y variable always goes on vertical axis by convention)

twoway (scatter tenure yearsed)



```
. reg tenure yearsed
```

Source	SS	df	MS	Number of obs =	5
Model	27.5625	1	27.5625	F(1, 3) =	4.48
Residual	18.4375	3	6.14583333	Prob > F =	0.1245
Total	46	4	11.5	R-squared =	0.5992
				Adj R-squared =	0.4656
				Root MSE =	2.4791

tenure	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
yearsed	1.3125	.6197698	2.12	0.124	-.6598841 3.284884
_cons	-11.0625	8.132929	-1.36	0.267	-36.94511 14.82011

The implied direction of causality in a regression is always from X (in this case years of education) to Y (in this case years of job tenure)

To obtain the residuals after a regression command in Stata type
predict uhat, resid

(this creates a new variable called uhat with the residuals for each observation)

the command
predict yhat

creates a new variable called yhat with the predicted values for each observation

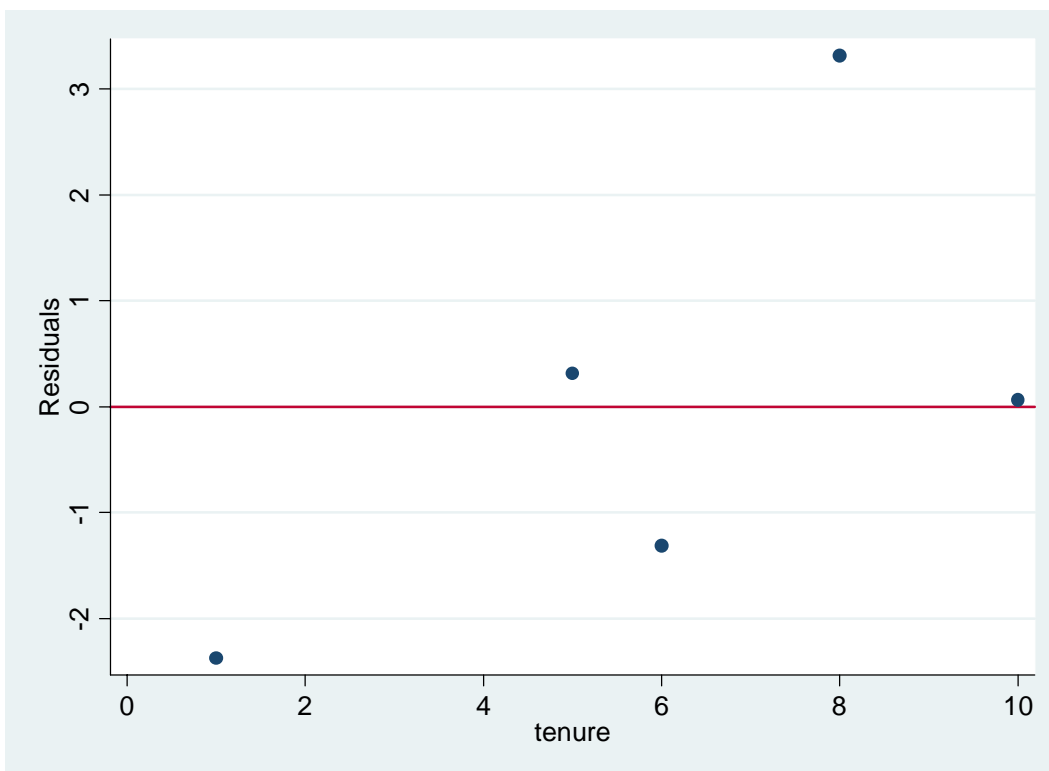
Can see this if type

list tenure yhat uhat

	tenure	yhat	uhat
1.	1	3.375	-2.375
2.	6	7.3125	-1.3125
3.	8	4.6875	3.3125
4.	10	9.9375	.0625
5.	5	4.6875	.3125

To graph the residuals for each y value

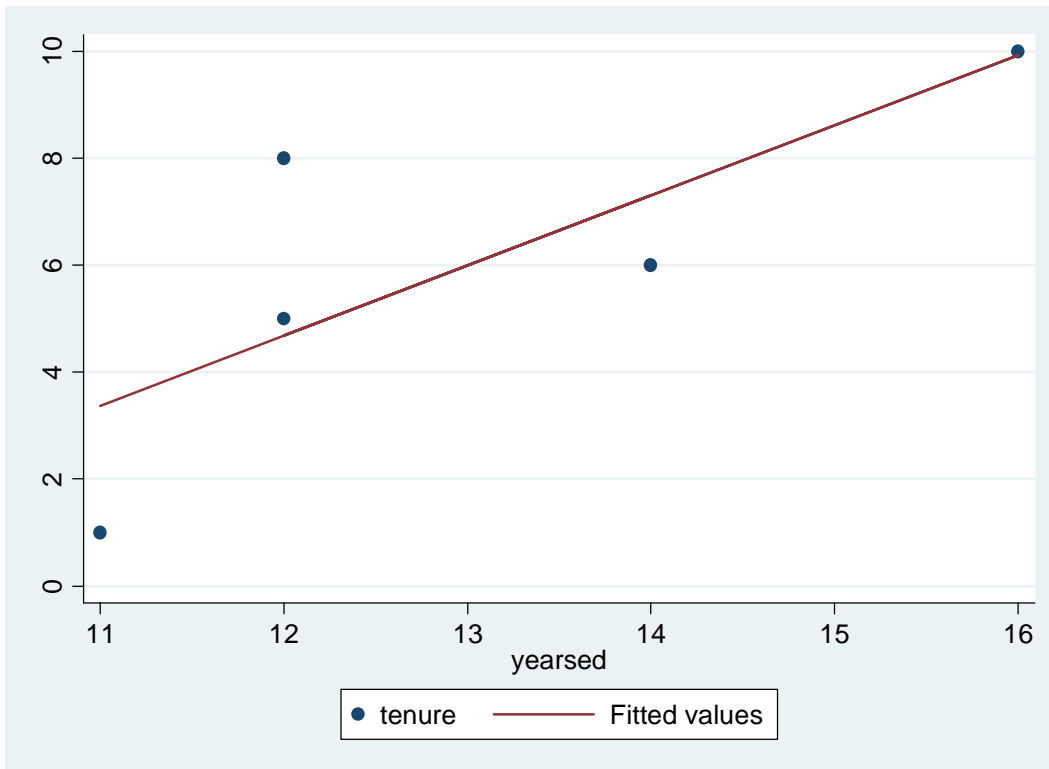
twoway (scatter uhat tenure, yline(0))



note how the residuals are scattered around zero (as they should be)

To graph the fitted line against the actual values

twoway (scatter tenure yearsed) (line yhat yearsed)



2.

$$\begin{aligned}
 \text{Cov}(\hat{Y}, e) &= \text{Cov}([b_1 + b_2 X], e) = \text{Cov}(b_1, e) + \text{Cov}(b_2 X, e) \\
 &= 0 + b_2 \text{Cov}(X, e) = b_2 \text{Cov}(X, [Y - b_1 - b_2 X]) \\
 &= b_2 [\text{Cov}(X, Y) - \text{Cov}(X, b_1) - \text{Cov}(X, b_2 X)] \\
 &= b_2 [\text{Cov}(X, Y) - b_2 \text{Cov}(X, X)] \\
 &= b_2 \left[\text{Cov}(X, Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \text{Var}(X) \right] = 0
 \end{aligned}$$

3. u psl

```
. reg emp gdp if country<24
```

Source	SS	df	MS	Number of obs = 23		
Model	14.8681911	1	14.8681911	F(1, 21) =	33.72	
Residual	9.26053015	21	.440977626	Prob > F =	0.0000	
Total	24.1287212	22	1.09676006	R-squared =	0.6162	
				Adj R-squared =	0.5979	
				Root MSE =	.66406	

empl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gdp	.4951757	.0852783	5.807	0.000	.3178298	.6725217
_cons	-.6091608	.2784317	-2.188	0.040	-1.188191	-.0301304

So a 1% point rise in the gdp rate is associated with a .495% point rise in employment growth.
 The constant suggests that if gdp growth were zero, then employment growth would be -.609 % points each year.

(Should also sketch fitted regression line. Can do this manually or in Stata)

Now adding in the U.S. and Japan gives

```
. reg emp gdp
```

Source	SS	df	MS	Number of obs = 25		
Model	14.5753023	1	14.5753023	F(1, 23)	=	33.10
Residual	10.1266731	23	.440290135	Prob > F	=	0.0000
-----				R-squared	=	0.5900
Total	24.7019754	24	1.02924898	Adj R-squared	=	0.5722
-----				Root MSE	=	.66354

empl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gdp	.489737	.0851184	5.754	0.000	.3136561	.6658179
_cons	-.5458912	.2740387	-1.992	0.058	-1.112784	.0210011

Comparing the estimates of the slope coefficients, can see the new slope is .489 Which indicates that the effect of a unit change in GDP on employment growth is marginally smaller, (.49 v. .495) given the extra data, as is the estimated slope coefficient, (-.55 v. -.61), so annual employment growth is now .55 percentage points a year at zero gdp.

Should not that R² now falls as a result of extra data, (from .62 to .59)
 Why?

Using formula and information in Table should be able to show that ratio of Explained sum of squares to total sum of squares has fallen, (14.868/24.128 in equation 1 v. 14.575/24.702 in equation 2)
 More information means, in this case, more variation in the dependent variable, but less in the explanatory variable. In other words the extra variation in employment growth provided by the 2 new countries can't be explained by the extra information in the gdp behaviour of the 2 countries.

Students might also note that the standard errors on both coefficients are largely unchanged, but that the t statistics fall, more so for the constant (See next problem set for more on t-stats)

Rescaling both variables, should not make any difference to slope estimates, but will change the value of the constant (proportionately)

```
. replace emp=emp/100
(25 real changes made)
```

```
. replace gdp=gdp/100
(25 real changes made)
```

```
. reg emp gdp
```

Source	SS	df	MS	Number of obs = 25		
Model	.00145753	1	.00145753	F(1, 23)	=	33.10
Residual	.001012667	23	.000044029	Prob > F	=	0.0000
-----				R-squared	=	0.5900

-----+-----				Adj R-squared = 0.5722		
Total		.002470197	24	.000102925	Root MSE = .00664	
-----+-----						
empl		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----						
gdp		.489737	.0851184	5.754	0.000	.3136561 .6658179
_cons		-.0054589	.0027404	-1.992	0.058	-.0111278 .00021
-----+-----						

Rescaling dependent variable only will change both slope and intercept estimates (proportionately by the amount of rescaling, in this case 1/100)

```
. replace gdp=gdp*100
(25 real changes made)
```

```
. reg emp gdp
```

-----+-----				Number of obs = 25	
Model		.00145753	1	.00145753	F(1, 23) = 33.10
Residual		.001012667	23	.000044029	Prob > F = 0.0000
-----+-----				R-squared = 0.5900	
Total		.002470197	24	.000102925	Adj R-squared = 0.5722
-----+-----				Root MSE = .00664	

-----+-----						
empl		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----						
gdp		.0048974	.0008512	5.754	0.000	.0031366 .0066582
_cons		-.0054589	.0027404	-1.992	0.058	-.0111278 .00021
-----+-----						

Note. Should now be able to show effects of rescaling only the explanatory variable.

Result of reverse regression is

```
. reg gdp empl
```

-----+-----				Number of obs = 25	
Model		35.8572977	1	35.8572977	F(1, 23) = 33.10
Residual		24.913042	23	1.08317574	Prob > F = 0.0000
-----+-----				R-squared = 0.5900	
Total		60.7703396	24	2.53209748	Adj R-squared = 0.5722
-----+-----				Root MSE = 1.0408	

-----+-----						
gdp		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----						
empl		1.204822	.2094033	5.754	0.000	.7716383 1.638006
_cons		1.81246	.2716574	6.672	0.000	1.250494 2.374426
-----+-----						

which suggests an alternative slope estimate of $1/1.204822 = .829$
which is a long way from .49 estimated above

4. Interpret the meaning of the OLS estimates of the constant and the slope in the following prediction equations.

$$\hat{Wage} = 5 + 1.2 * Age$$

(where wage is measured in £ an hour and age is measured in years)

so OLS estimate of slope = $dWage/dAge = 1.2$
and 1 more year increases wages by £1.20 an hour

OLS estimate of constant suggests that if age were zero than wages would be £5 an hour (sometimes the will make sense, other times, like here, the intercept will have no real world interpretation)

$$\hat{Consumption} = 3,000 + 0.82 * Income$$

(where annual consumption and income levels are measured in £000)

so OLS estimate of slope = $dConsumption/dIncome = 0.82$
and £1000 increase in income raises annual consumption by $0.82 * £1000$ ie £820 a year

The OLS estimate of the constant suggests that if income were ever zero then consumption would be £3000 a year (perhaps people use savings in order to consume if income is zero)

$$\hat{GDP} = -5,000 + 1000 * Population$$

(where GDP is measured in \$ and population is measured in millions)

so OLS estimate of slope = $dGDP/dPopulation = 1000$
and 1 million more people increases GDP by \$1000

and estimate of constant suggests GDP would be -\$5000 if population were ever zero

$$\hat{Weight} = -210 + 0.51 * Height$$

(where weight is measured in kilograms and height in centimetres)

so OLS estimate of slope = $dWeight/dHeight = 0.51$
and 1 additional centimeter of height increases weight by 0.51 kilograms

and the constant gives a notional weight of -210 Kg at height zero

5.

a) True b) True c) True d) False e) False f) False

6. Use $\hat{\beta} = Cov(X,Y)/Var(X)$

Rescaling the variables is the same as mutiplying the variable by a constant value.

If height is measured in centimeters rather than meters then the new right hand side X variable becomes =100X

Using rules on variance and covariance (see problem set 1)

$$Cov(ax, Y) = aCov(X, Y)$$

$$Var(ax) = a^2Var(X)$$

So the new estimated coefficient (when height is measured in centimeters),

$$\begin{aligned}\tilde{\beta} &= \text{Cov}(aX, Y) / \text{Var}(aX) = a \text{Cov}(X, Y) / a^2 \text{Var}(X) = \text{Cov}(X, Y) / a \text{Var}(X) \\ &= \hat{\beta} / 100 = 0.006\end{aligned}$$

ie if the right hand side variable is multiplied by a then the OLS estimate of the slope is multiplied by $1/a$ (and the constant is unchanged) – Check this using the consumption function data on the P drive. This makes sense the relationship between the variables is unchanged only the units of measurement are different. Now $d\text{Weight}/d\text{Height}$ measures the effect of a **1cm** increase in height on weight in kilograms, rather than the effect of a **1m** increase in height