

OLS Standard Errors and Heteroskedasticity

Read in the data set cluster.dta. Consider a regression of whether an individual is employed conditional on age, gender and the average wage in the region, where there are 10 regions in the data set.

```
. reg emp age sex mregw if ures<11
```

Source	SS	df	MS			
Model	37.4895549	3	12.4965183	Number of obs = 11261		
Residual	2325.5988	11257	.206591348	F(3, 11257) = 60.49		
				Prob > F = 0.0000		
				R-squared = 0.0159		
				Adj R-squared = 0.0156		
Total	2363.08836	11260	.209865751	Root MSE = .45452		

emp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0005636	.0003215	1.75	0.080	-.0000665	.0011937
sex	-.1104338	.0085695	-12.89	0.000	-.1272316	-.093636
mregw	-.0082664	.0025087	-3.30	0.001	-.013184	-.0033488
_cons	.9377022	.0344903	27.19	0.000	.8700951	1.005309

Since this is a multiple regression, inspection of the residuals to check for heteroskedasticity is harder. Hence use Breusch-Pagan and White tests

```
bpagan age sex mregw
```

Breusch-Pagan LM statistic: 68.4366 Chi-sq(3) P-value = 9.2e-15

```
whitetst
```

White's general test statistic : 1043.962 Chi-sq(8) P-value = 5.e-220

Both tests reject null of homoskedasticity. (Remember Breusch-Pagan test is more flexible since can chose variables thought to influence heteroskedasticity)

As a result of the above, decide to correct the OLS standard errors to be robust to unknown form of heteroskedasticity

```
reg emp age sex mregw if ures<11, robust
```

emp	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0005636	.0003555	1.59	0.113	-.0001332	.0012604
sex	-.1104338	.0085453	-12.92	0.000	-.1271841	-.0936834
mregw	-.0082664	.0025886	-3.19	0.001	-.0133406	-.0031923
_cons	.9377022	.0354621	26.44	0.000	.8681903	1.007214

Linear regression
 Number of obs = 11261
 F(3, 11257) = 60.55
 Prob > F = 0.0000
 R-squared = 0.0159
 Root MSE = .45452

As a result standard error on age falls and t value rises, though levels of significance don't change much.

However since there are only 10 points of variation in the regional wage variable, use the clustering correction.

```
reg emp age sex mregw if ures<11, cluster(ures)
```

Linear regression

Number of obs = 11261
 F(3, 9) = 46.63
 Prob > F = 0.0000
 R-squared = 0.0159
 Root MSE = .45452

(Std. Err. adjusted for 10 clusters in uresmc)

emp	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0005636	.0009297	0.61	0.559	-.0015396	.0026668
sex	-.1104338	.0093796	-11.77	0.000	-.131652	-.0892155
mregw	-.0082664	.008623	-0.96	0.363	-.0277729	.0112401
_cons	.9377022	.0987054	9.50	0.000	.7144152	1.160989

While the standard errors on age and sex change (there is some regional variation in these variables), the standard error on the regional wage variable doubles, so that the t value now indicates no significant effect of the regional wage. Changes of this size are typical when clustering is heavily concentrated in this way.