

Grouped Data and Weighted Least Squares (GLS)

The data set reg.dta has information on wages and employment rates for 20 UK regions

The sample population of each region with non-missing wage observations is

```
list ures npop
```

	+-----+	
	uresmc	npop
	+-----+	
1.	tyne & w	1628
2.	rest of	2930
3.	south yo	1937
4.	west yor	3136
5.	rest of	2435
	+-----+	
6.	east mid	5899
7.	east ang	3137
8.	inner lo	3537
9.	outer lo	6144
10.	rest of	15950
	+-----+	
11.	south we	6945
12.	west mid	3592
13.	rest of	3935
14.	greater	3472
15.	merseysi	1869
	+-----+	
16.	rest of	3254
17.	wales	4156
18.	strathcl	3463
19.	rest of	4542
20.	northern	3615

In this case the variance of the residual in each region is given by $\text{Var}u_i = \sigma_u^2 \Omega = \sigma_u^2 / N_r$ where N_r is the population in each region. Hence

$$\Omega = \begin{bmatrix} 1/N_1 & & 0 \\ & \dots & \\ 0 & & 1/N_{20} \end{bmatrix} \quad \text{and} \quad \Omega^{-1} = \begin{bmatrix} N_1 & & 0 \\ & \dots & \\ 0 & & N_{20} \end{bmatrix}$$

Given $\hat{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y = \sum_{r=1}^{20} \Omega_r x_r x_r'$

then the GLS transformation requires multiplying the original observations by the **square root** of the population in each region

```
g srtnpop=sqrt(npop)
g sregw=regw*srtn
g srege=rege*srtn
g srega=rega*srtn
g sregh=regh*srtn
g glscons=1*srtn          /* remember the constant must also be transformed */
```

The OLS estimate is

```
. reg regemp regw rega
```

Source	SS	df	MS			
Model	.035492788	2	.017746394	Number of obs =	20	
Residual	.015978718	17	.000939925	F(2, 17) =	18.88	
Total	.051471506	19	.002709027	Prob > F =	0.0000	
				R-squared =	0.6896	
				Adj R-squared =	0.6530	
				Root MSE =	.03066	

	regemp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	regw	.0005167	.000177	2.92	0.010	.0001431	.0008902
	regage	.0552779	.009017	6.13	0.000	.0362538	.0743021
	_cons	-1.584658	.3786123	-4.19	0.001	-2.383461	-.7858564

The weighted least squares GLS estimator can be obtained in stata by using the "analytic weight" option and specifying the variable to weight by (in this case the population in each region)

```
. reg regemp regw rega [aw=npop]
(sum of wgt is 8.5576e+04)
```

Source	SS	df	MS			
Model	.040384696	2	.020192348	Number of obs =	20	
Residual	.01112938	17	.000654669	F(2, 17) =	30.84	
Total	.051514076	19	.002711267	Prob > F =	0.0000	
				R-squared =	0.7840	
				Adj R-squared =	0.7585	
				Root MSE =	.02559	

	regemp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	regw	.0005708	.0001393	4.10	0.001	.0002769	.0008646
	regage	.0591121	.0075416	7.84	0.000	.0432006	.0750235
	_cons	-1.744273	.3139756	-5.56	0.000	-2.406704	-1.081842

or manually

```
. reg srege sregw srega glscons, nocons
```

Source	SS	df	MS			
Model	44127.782	3	14709.2607	Number of obs =	20	
Residual	47.6203744	17	2.8011985	F(3, 17) =	5251.06	
Total	44175.4024	20	2208.77012	Prob > F =	0.0000	
				R-squared =	0.9989	
				Adj R-squared =	0.9987	
				Root MSE =	1.6737	

	srege	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	sregw	.0005708	.0001393	4.10	0.001	.0002769	.0008646
	srega	.0591121	.0075416	7.84	0.000	.0432006	.0750235
	glscons	-1.744273	.3139755	-5.56	0.000	-2.406703	-1.081842

(The "noconstant" option suppresses the original constant allowing replacement by the transformed constant)