Weighted Least-Squares Regression

For regressions with cross-section data (where the subscript "i" denotes a particular individual or firm at a point in time), it is usually safe to assume the errors are uncorrelated, but often their variances are not constant across individuals. This is known as the problem of heteroskedasticity (for "unequal scatter"); the usual assumption of constant error variance is referred to as homoskedasticity. Although the mean of the dependent variable might be a linear function of the regressors, the variance of the error terms might also depend on those same regressors, so that the observations might "fan out" in a scatter diagram.

Approaches to Dealing with Heteroskedasticity

. For known heteroskedasticity (e.g., grouped data with known group sizes), use weighted least squares (WLS) to obtain efficient unbiased estimates;

. Test for heteroskedasticity of a special form using a squared residual regression;

. Estimate the unknown heteroskedasticity parameters using this squared residual regression, then use the estimated variances in the WLS formula to get efficient
estimates of regression coefficients (known as feasible WLS); or

. Stick with the (inefficient) least squares estimators, but get estimates of standard errors which are correct under arbitrary heteroskedasticity.

In this procedure, the "residualization" method is used to obtain weights that will reduce the effect of heteroskedastic values. The method consists of four stages:

Step 1. Perform an Ordinary Least Squares (OLS) regression and obtain the residuals and squared residuals where the residual is the difference between the observed dependent variable and the predicted dependent variable value for each case.

Step 2. Regress the values of the squared residuals on the independent variables using OLS. The F test for the model is an indication of heteroskedasticity in the data.

Step 3. Obtain the reciprocal of the square root of the absolute squared residuals. These weights are then multiplied times all of the variables of the regression model.

Step 4. Obtain the OLS regression of the weighted dependent variable on the weighted independent variables. One can obtain the regression through the origin. If elected, each variable's values are converted to deviations from their respective mean before the OLS analysis is performed.

As an alternative, the user may use weights he or she has derived. These should be similar to the reciprocal values obtained in step 3 above. When these weights are used, they are multiplied times the values of each variable and step 4 above is completed.

Shown below is the dialog box for the Weighted Least Squares Analysis and an analysis of the cansas.OS4 data file.

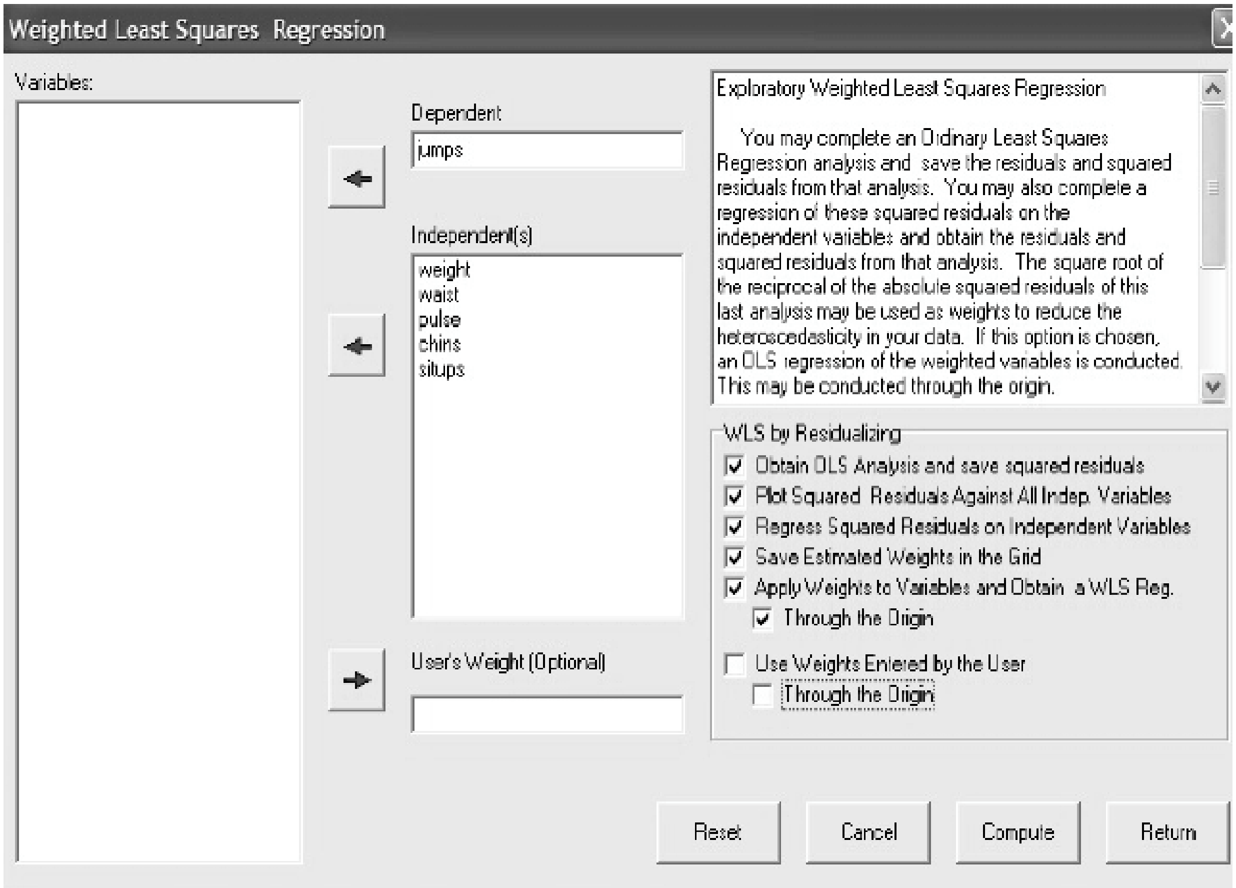


Figure 1 Weighted Least Squares Regression Dialog

OLS REGRESSION RESULTS

Means

Variables	weight	waist	pulse	chins	situps	jumps
	178.600	35.400	56.100	9.450	145.550	70.300

Standard Deviations

Variables	weight	waist	pulse	chins	situps	jumps
	24.691	3.202	7.210	5.286	62.567	51.277

No. of valid cases = 20

CORRELATION MATRIX

VARIABLE	weight	waist	pulse	chins	situps	jumps
weight	1.000	0.870	-0.366	-0.390	-0.493	-0.226
waist	0.870	1.000	-0.353	-0.552	-0.646	-0.191
pulse	-0.366	-0.353	1.000	0.151	0.225	0.035

chins	-0.390	-0.552	0.151	1.000	0.696	0.496
situps	-0.493	-0.646	0.225	0.696	1.000	0.669
jumps	-0.226	-0.191	0.035	0.496	0.669	1.000

Dependent variable: jumps

Variable	Beta	B	Std.Err.	t	Prob.>t	VIF	TOL
weight	-0.588	-1.221	0.704	-1.734	0.105	4.424	0.226
waist	0.982	15.718	6.246	2.517	0.025	5.857	0.171
pulse	-0.064	-0.453	1.236	-0.366	0.720	1.164	0.859
chins	0.201	1.947	2.243	0.868	0.400	2.059	0.486
situps	0.888	0.728	0.205	3.546	0.003	2.413	0.414
Intercept	0.000	-366.967	183.214	-2.003	0.065		

SOURCE	DF	SS	MS	F	Prob.>F
Regression	5	31793.741	6358.748	4.901	0.0084
Residual	14	18164.459	1297.461		
Total	19	49958.200			

R2 = 0.6364, F = 4.90, D.F. = 5 14, Prob>F = 0.0084
Adjusted R2 = 0.5066

Standard Error of Estimate = 36.02

REGRESSION OF SQUARED RESIDUALS ON INDEPENDENT VARIABLES

Means

Variables	weight	waist	pulse	chins	situps	ResidSqr
	178.600	35.400	56.100	9.450	145.550	908.196

Standard Deviations

Variables	weight	waist	pulse	chins	situps	ResidSqr
	24.691	3.202	7.210	5.286	62.567	2086.828

No. of valid cases = 20

CORRELATION MATRIX

VARIABLE	weight	waist	pulse	chins	situps	ResidSqr
weight	1.000	0.870	-0.366	-0.390	-0.493	-0.297
waist	0.870	1.000	-0.353	-0.552	-0.646	-0.211
pulse	-0.366	-0.353	1.000	0.151	0.225	-0.049
chins	-0.390	-0.552	0.151	1.000	0.696	0.441
situps	-0.493	-0.646	0.225	0.696	1.000	0.478
ResidSqr	-0.297	-0.211	-0.049	0.441	0.478	1.000

Dependent variable: ResidSqr

Variable	Beta	B	Std.Err.	t	Prob.>t	VIF	TOL
weight	-0.768	-64.916	36.077	-1.799	0.094	4.424	0.226
waist	0.887	578.259	320.075	1.807	0.092	5.857	0.171
pulse	-0.175	-50.564	63.367	-0.798	0.438	1.164	0.859
chins	0.316	124.826	114.955	1.086	0.296	2.059	0.486
situps	0.491	16.375	10.515	1.557	0.142	2.413	0.414
Intercept	0.000	-8694.402	9389.303	-0.926	0.370		

SOURCE	DF	SS	MS	F	Prob.>F
Regression	5	35036253.363	7007250.673	2.056	0.1323
Residual	14	47705927.542	3407566.253		
Total	19	82742180.905			

R2 = 0.4234, F = 2.06, D.F. = 5 14, Prob>F = 0.1323

Adjusted R2 = 0.2175

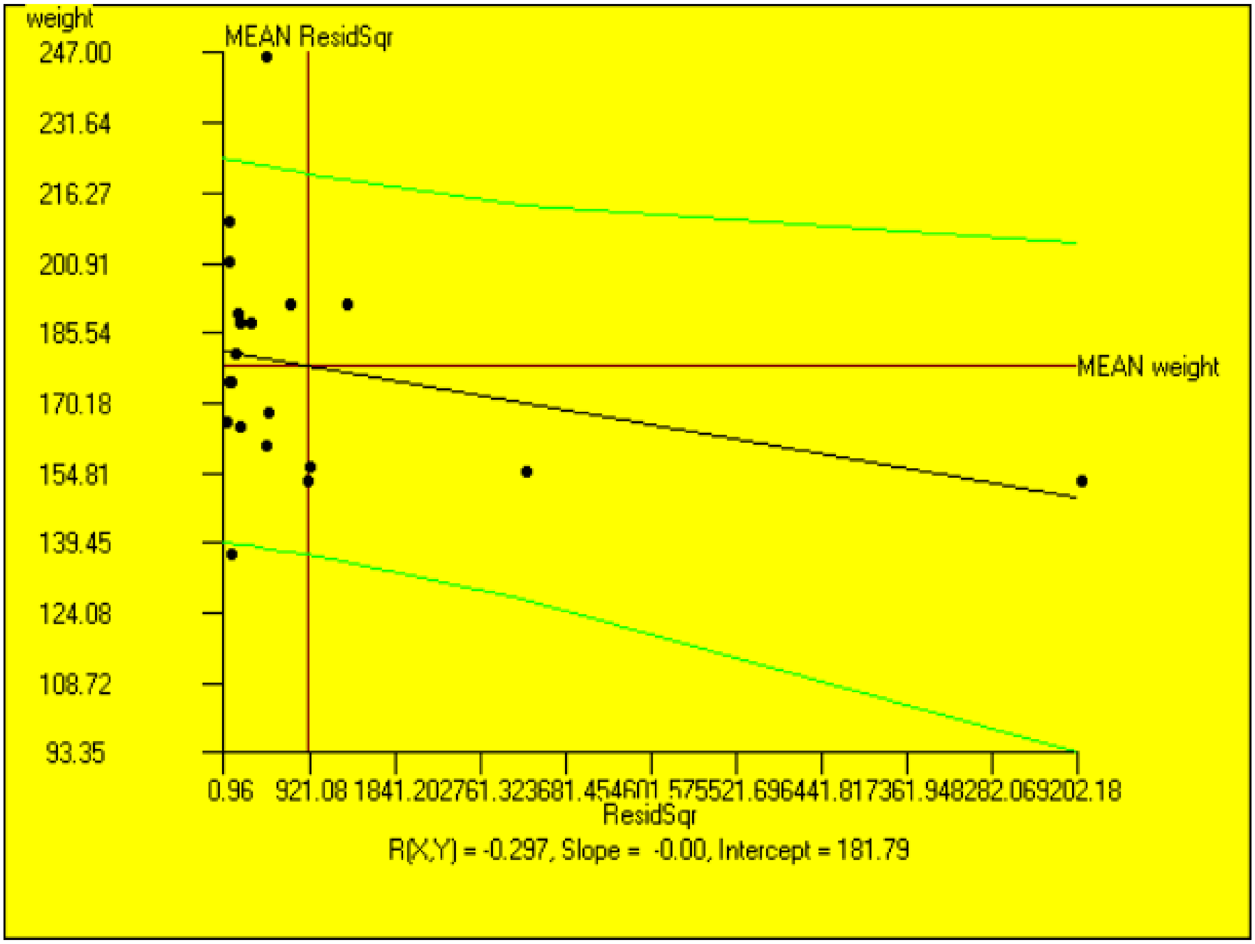
Standard Error of Estimate = 1845.96

X versus Y Plot

X = ResidSqr, Y = weight from file: C:\Documents and Settings\Owner\My Documents\Projects\Clanguage\OpenStat4\cansaswls.OS4

Variable	Mean	Variance	Std.Dev.
ResidSqr	908.20	4354851.63	2086.83
weight	178.60	609.62	24.69

Correlation = -0.2973, Slope = -0.00, Intercept = 181.79
Standard Error of Estimate = 23.57
Number of good cases = 20



WLS REGRESSION RESULTS
Means

Variables	weight	waist	pulse	chins	situps	jumps
	-0.000	0.000	-0.000	0.000	-0.000	0.000

Standard Deviations

Variables	weight	waist	pulse	chins	situps	jumps
	7.774	1.685	2.816	0.157	3.729	1.525

No. of valid cases = 20

CORRELATION MATRIX

VARIABLE	weight	waist	pulse	chins	situps	jumps
weight	1.000	0.994	0.936	0.442	0.742	0.697
waist	0.994	1.000	0.965	0.446	0.783	0.729
pulse	0.936	0.965	1.000	0.468	0.889	0.769
chins	0.442	0.446	0.468	1.000	0.395	0.119
situps	0.742	0.783	0.889	0.395	1.000	0.797
jumps	0.697	0.729	0.769	0.119	0.797	1.000

Dependent variable: jumps

Variable	Beta	B	Std.Err.	t	Prob.>t	VIF	TOL
weight	-2.281	-0.448	0.414	-1.082	0.298	253.984	0.004
waist	3.772	3.415	2.736	1.248	0.232	521.557	0.002
pulse	-1.409	-0.763	0.737	-1.035	0.318	105.841	0.009
chins	-0.246	-2.389	1.498	-1.594	0.133	1.363	0.734
situps	0.887	0.363	0.165	2.202	0.045	9.258	0.108
Intercept	0.000	-0.000	0.197	-0.000	1.000		

SOURCE	DF	SS	MS	F	Prob.>F
Regression	5	33.376	6.675	8.624	0.0007
Residual	14	10.837	0.774		
Total	19	44.212			

R2 = 0.7549, F = 8.62, D.F. = 5 14, Prob>F = 0.0007
Adjusted R2 = 0.6674

Standard Error of Estimate = 0.88