Regression weighting for a discrete predictor using means

Stephen Senn

Set up

We suppose that we have an exposure variable $X_i, i = 1 \cdots N$ that can only take on certain discrete values say, $D_j, j = 1 \cdots k, k \ll N$. In other words, the number of subjects greatly exceeds the number of possible exposure categories. The data have been summarised in the form of exposure D_j , the mean outcome, $\bar{Y}_j$, for all subjects with that exposure and the numbers of subjects, n_j who had that exposure.

Solution

The ordinary least squares estimate of the slope in terms of the original data is

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{N} w_i Y_i$$

where

$$w_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

is a weight. As is well known the OLS estimator is a weighted sum of the outcome values.

Now without loss of generality, suppose the values sorted so that all subject with the same exposure are grouped together. It is obvious that we can express the overall slope as

$$\hat{\beta} = \sum_{j=1}^{k} \hat{\beta}_j$$

where $\hat{\beta}_j$ is the contribution to the total weighted sum for the subjects with exposure j .

For a given exposure group, j , we may now write the weights as

$$w_j = \frac{\left(D_j - \bar{X}\right)}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \ .$$

Note that for any subjects in the same exposure group, these weights are constant.

Each outcome value in the given exposure group is multiplied by this same constant so the contribution from exposure group j is w_j times the sum of the outcomes or, equivalently, the mean multiplied by the number of subjects. Thus we have

$$\hat{\beta}_j = \frac{n_j\left(D_j - \bar{X}\right)\bar{Y}_j}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Note also, that we have

$$\bar{X} = \frac{\sum_{j=1}^{k} n_j D_j}{N}$$

and

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{j=1}^{k} n_j (D_j - \bar{X})^2$$

Intercept

The intercept may be calculated by calculating weighted means for the predictor and the outcome. Thus we can calculate

$$\bar{D} = \bar{X} = \frac{\sum_{j=1}^{k} n_j D_j}{\sum_{j=1}^{k} n_j}, \quad \bar{Y} = \frac{\sum_{j=1}^{k} n_j \bar{Y}_j}{\sum_{j=1}^{k} n_j}$$

and hence

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

for the intercept.

Weighted least squares

These formulae imply that the slope and intercept can be estimated from any regression package capable of carrying out weighted regression by using group predictor and outcome means as raw input with numbers of observations in each group as weights.

Standard Errors

The same is not true, however, of standard errors. This is because the overall ability of the model to predict depends not just on its ability to predict group means but also individual values. Although the original X_i values are recoverable from the predictor values D_j and the number of subjects n_j having the given value, the individual Y_i values cannot be recovered from the group outcome means $\overline{Y}_j$. However, if the group variances are available, then the value of the standard error of the least squares estimate using original values can be recovered using the following steps.

1. Obtain the between groups sum of squares, SSB from the weighted regression
2. Calculate the within group sums of squares as $SSW = \sum_{j=1}^{k}\left[\left(n_j - 1\right)\mathrm{var}\left(Y_j\right)\right]$
3. Calculate the total sums of squares as $SST = SSB + SSW$
4. Estimate the variance of the disturbance terms in the model as $\hat{\sigma}^2 = SST/\left(N-2\right)$
5. Estimate the standard error of $\hat{\beta}$ as $SE\left(\hat{\beta}\right) = \sqrt{\hat{\sigma}^2 \Big/ \sum_{i=1}^{N}(X_i - \overline{X})^2}$

Weighting in a meta-analysis

If it is only possible to estimate the standard error by weighted regression it will be based on few degrees of freedom and extremely unstable[1]. Weighting using the inverse variance method will not be appropriate and it will often be better to use $\sum_{i=1}^{n}(X_i - \overline{X})^2$ as a weight.

Stephen Senn, 26 February 2018, updated 27 February 2018

Reference

[1] Senn, S.J. 2000 The many modes of meta. *Drug Information Journal* **34**, 535-549.

Example

Genstat Nineteenth Edition
Genstat Procedure Library Release PL27.1

Illustration of grouped regression

Print original data

X	Exposure	Y
1	1	3.041
1	1	5.135
1	1	3.158
1	1	3.047
1	1	1.975
1	1	4.546
1	1	5.128
2	2	8.098
2	2	8.897
2	2	7.726
2	2	6.394
2	2	8.588
2	2	7.879
3	3	13.112
3	3	10.595
3	3	10.121
3	3	9.649
3	3	11.565
4	4	16.315
4	4	13.789

Calculation using original individual exposure data

Y	Predicted	Residuals	Residuals^2
3.041	3.863	-0.822	0.676
5.135	3.863	1.272	1.617
3.158	3.863	-0.705	0.497
3.047	3.863	-0.816	0.666
1.975	3.863	-1.889	3.566
4.546	3.863	0.683	0.466
5.128	3.863	1.264	1.599
8.098	7.567	0.530	0.281
8.897	7.567	1.329	1.767
7.726	7.567	0.158	0.025
6.394	7.567	-1.173	1.376
8.588	7.567	1.021	1.041
7.879	7.567	0.312	0.097
13.112	11.272	1.840	3.387
10.595	11.272	-0.677	0.458
10.121	11.272	-1.151	1.324
9.649	11.272	-1.623	2.633
11.565	11.272	0.294	0.086
16.315	14.976	1.339	1.793
13.789	14.976	-1.187	1.409

Intercept	Slope calculated from original data
0.1590	3.704

CSS_X, corrected sum of squares for X	Sum of squared residuals
19.80	24.77

Estimate of σ^2	Standard error of slope
1.376	0.2636

Estimate of σ^2 is (Sum of squared residuals)/(20-2)

Estimate of slope standard error is $\sqrt{(\sigma^2/CSS_X)}$

Using standard regression package on original data

Regression analysis

Estimates of parameters

Parameter	estimate	s.e.	t(18)
Constant	0.159	0.613	0.26
X	3.704	0.264	14.05

Calculation using grouped data

mean exposure	mean outcome	corrected sum of squares of X
2.100	7.938	19.80

observations	exposure	Exposure Difference	weight	mean outcome	product
7	1	-1.100	-0.3889	3.719	-1.446
6	2	-0.100	-0.0303	7.930	-0.240
5	3	0.900	0.2273	11.008	2.502
2	4	1.900	0.1919	15.052	2.889

Grouped intercept	grouped slope
0.1590	3.704

observations	mean outcome	Group predicted	Group residuual	Group residual^2
7	3.719	3.863	-0.1447	0.02094
6	7.930	7.567	0.3629	0.13172
5	11.008	11.272	-0.2633	0.06933
2	15.052	14.976	0.0759	0.00576

n x Group residual^2
0.1466
0.7903
0.3467
0.0115

between SS	within SS
1.295	23.47

Within SS calculated by taking the variance group by group x DF and adding

Total SS
24.77

Total SS calculated by adding between and within SS

Illustration using weighted regression

Note: The standard errors will not be the same as they would be using original data because the within group variation is ignored

Regression analysis

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.
Regression	1	271.678	271.6778	419.56
Residual	2	1.295	0.6475	
Total	3	272.973	90.9910	

Percentage variance accounted for 99.3
Standard error of observations is estimated to be 0.805.

Estimates of parameters

Parameter	estimate	s.e.	t(2)
Constant	0.159	0.420	0.38
exposure	3.704	0.181	20.48