diagonal matrix with elements

$$W_{ii} = \begin{cases} q/f_{\text{errors}}(0) & \text{if } r > 0 \\ (1-q)/f_{\text{errors}}(0) & \text{if } r < 0 \\ 0 & \text{otherwise} \end{cases}$$

and $\mathbf{R}_2$ is the design matrix $\mathbf{X}'\mathbf{X}$. This is derived from formula 3.11 in Koenker and Bassett, although their notation is much different. $f_{\text{errors}}()$ refers to the density of the true residuals. There are many things that Koenker and Bassett leave unspecified, including how one should obtain a density estimate for the errors in real data. It is at this point that we offer our contribution.

We first sort the residuals and locate the observation in the residuals corresponding to the quantile in question, taking into account weights if they are applied. We then calculate w_n, the square root of the sum of the weights. Unweighted data is equivalent to weighted data where each observation has weight 1, resulting in $w_n = \sqrt{n}$. For analytically weighted data, the weights are rescaled so that the sum of the weights is the sum of the observations, resulting in $\sqrt{n}$ again. For frequency weighted data, w_n literally is the square of the sum of the weights.

We locate the closest observation in each direction such that the sum of weights for all closer observations is w_n. If we run off the end of the dataset, we stop. We calculate w_s, the sum of weights for all observations in this middle space. Typically, w_s is slightly greater than w_n.

The residuals obtained after quantile regression have the property that if there are k parameters, then exactly k of them must be zero. Thus, we calculate an adjusted weight $w_a = w_s - k$. The density estimate is the distance spanned by these observations divided by w_a. Because the distance spanned by this mechanism converges toward zero, this estimate of the density converges in probability to the true density.

References

Gould, W. 1992. Quantile regression and bootstrapped standard errors. *Stata Technical Bulletin* 9: 19–21.

Koenker, R. and G. Bassett, Jr. 1982. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* 50: 43–61.

Rogers, W. H. 1992. Quantile regression standard errors. *Stata Technical Bulletin* 9: 16–19.

sg17 Regression standard errors in clustered samples

William Rogers, CRC, FAX 310-393-7551

Stata's `hreg`, `hlogit` and `hprobit` commands estimate regression, maximum-likelihood logit, and maximum-likelihood probit models based on Huber's (1967) formula for individual-level data and they produce consistent standard errors even if there is heteroscedasticity, clustered sampling, or the data is weighted. The description of this in [5s] hreg might lead one to believe that Huber originally considered clustered data, but that is not true. I developed this approach to deal with cluster sampling problems in the RAND Health Insurance Experiment in the early 1980s (Rogers 1983; Rogers and Hanley 1982; Brook, et al. 1983). What is true is that with one simple assumption, the framework proposed by Huber can be applied to produce the answer we propose. That assumption is that the clusters are drawn as a simple random sample from some population. The observations must be obtained within each cluster by some repeatable procedure.

Ordinary linear regression applied to the observations of a cluster is a nonstandard maximum-likelihood estimate; that is, a maximum of the "wrong" likelihood, given this sampling procedure. This is an important special case of the more general problem that Huber's article addresses.

The special case can be obtained by recognizing that a cluster can play the same role as an observation. For instance, the Huber regularity conditions require that the importance of any one observation vanishes as the number of observations becomes infinite. Huber's method is not good when there are only a few observations. In this special case, the Huber regularity conditions require that the importance of any one *cluster* vanishes as the number of clusters (and therefore observations) becomes infinite. Thus, Huber's reinterpreted method is not good when there are only a few clusters.

To apply Huber' formula directly, one would want to calculate a score value and a Hessian value for each cluster. This is described in [5s] huber. Although elegant, this application of Huber's result does not provide much insight into why the idea works. For the case of linear regression, the matrix algebra provides more insight.

Let p be the number of parameters and n the number of observations. Let $\mathbf{X}$ be the $n \times p$ design matrix and $\mathbf{y}$ be the $n \times 1$ vector of dependent values. The ordinary linear regression estimates are $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The variance of this estimate is

$$\text{var}(\mathbf{b}) = \text{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \text{E}\mathbf{y})(\mathbf{y} - \text{E}\mathbf{y})'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

If we adopt the usual practice of replacing unknown errors with residuals, the inner matrix is an $n \times n$ rank 1 matrix, which is not very helpful. The original solution to this problem is to assume that the $\mathbf{X}$ matrix is fixed and move the expectation inside, and take advantage of the independent and identically distributed assumption to assert that $\mathrm{E}(\mathbf{y} - \mathrm{E}\mathbf{y})(\mathbf{y} - \mathrm{E}\mathbf{y})' = \sigma^2\mathbf{I}$. All of the off-diagonal terms are zero, and all of the diagonal terms are the same. The estimate of σ^2 is obtained by substituting residuals for the errors $(\mathbf{y} - \mathrm{E}\mathbf{y})$. After reduction, the familiar variance estimate $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ is obtained.

In the revised solution, we do not assume that the diagonal terms are identical (White 1980). Also, we do not assume off-diagonal terms are zero unless they come from different clusters. Observations from different clusters are independent, so their off-diagonal elements must be zero. We simply let all these nonzero terms be represented by the appropriate products of the residuals.

Ordinarily, estimating n parameters, or even more with clustering, would be a bad idea. However, with pre- and post-multiplication by $\mathbf{X}$, a convergent averaging effect is obtained provided that no cluster is too large.

If weights are present, these weights appear in the equation and are treated as part of the observations. The variance estimate now becomes:

$$\mathrm{var}(\mathbf{b}) = \mathrm{E}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(\mathbf{y} - \mathrm{E}\mathbf{y})(\mathbf{y} - \mathrm{E}\mathbf{y})'\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$$

Since linear regression is not the maximum-likelihood answer, most statisticians would presume that it does not give an answer we would want. However, it is worth pointing out that the "wrong" answer given by linear regression is the answer that would be given if the entire population of clusters were sampled in the manner prescribed. In some applications this is the desired answer, and other answers converge to something else. In each case, the user must decide if the linear regression answer is wanted or not, on a theoretical basis. For example, if we sample families and then take one family member (without weighting), family members in large families will be undersampled.

Two advantages of this framework over other approaches to the cluster sampled problem are (1) that the nature of the within-cluster dependence does not have to be specified in any way, and (2) familiar estimation methods can be used. Since the method can be thought of as "correcting" linear regression, the user is free to conduct other types of sensitivity analysis in parallel. For example, he might also consider a sample selection model using the linear regression results as a common point of comparison.

Although the mathematics guarantees that the large sample behavior of this estimate will be good, what about small-sample behavior? A few Monte-Carlo experiments will give a good idea of what is going on. Simple experiments will suffice since Huber covariance estimates respond to affine transformations of the $\mathbf{X}$ matrix or $\mathbf{y}$ observation vector just as regular covariance estimates do.

Experiment 1

First, we verify that in large samples the answers obtained by the Huber algorithm is okay. We draw 2,500 observations clustered in groups of 5 as follows:

```
. clear
. set obs 2500
. gen x = (_n-1250.5)/1249.5
. gen y = invnorm(uniform())
. gen u = uniform()
. gen g = int((_n-1)/5)
```

We then run 1,000 examples of this and examine the collective results.

The known covariance matrix for the ordinary linear regression coefficients is

$$\begin{pmatrix} 0.0004000 & 0 \\ 0 & 0.001199 \end{pmatrix}$$

For the standard regression estimates, these covariances are obtained up to a multiplier that is distributed $\chi^2(2498)/2498$, which has expectation 1 and variance $2/2498$.

We will look at two Huber estimates. The first Huber estimate assumes no clustering and is equivalent to White's method. The second Huber estimate assumes clustering in 500 clusters of 5 observations each. Each cluster contains nearby values of x.

There are two things we would want to know. First, do these variance estimation procedures estimate the right variance on the average, and second, how well do the estimates reflect known population behavior?

		Huber/White method	
	Usual formula	Unclustered	Clustered
Average estimated variance of the coefficient			
var(_cons) $\times 10^4$	4.000	3.995	3.979
var(_b[x]) $\times 10^4$	11.99	11.98	11.93
correlation	0.	0.	0.
RMS error of the variance estimate			
var(_cons) $\times 10^8$	32	115.	2439.
var(_b[x]) $\times 10^8$	96	4973.	9730.
correlation	0	.028	.063
Percent of cases marked as significant			
var(_cons)	5.0	4.3	4.3
var(_b[x])	5.0	4.7	4.9

All three methods produce essentially the same variance estimates. There is no cost here for added robustness. The unclustered and clustered Huber estimates of the variance are more variable, but it does not matter. Asymptotics have taken hold.

Experiment 2

Next, we verify the desirable properties of the Huber estimate for clustered data in large samples. We draw 2,500 observations clustered in groups of 5 as follows:

```
. clear
. set obs 2500
. gen x = (_n-1250.5)/1249.5
. gen y = invnorm(uniform())
. gen yy = invnorm(uniform())
. gen u = uniform()
. gen g = int((_n-1)/5)
. sort g
. qui by g: replace y = sqrt(.8)*y + sqrt(.2)*yy[_N] if _n < _N
```

We then run 1,000 examples of this and examine the collective results. The intracluster correlation is 0.2, and the group size is 5, so the design effect DEFF (see [5s] deff) for this problem is 1.8, meaning that the F-ratios of the standard problem are too high by a multiple of 1.8.

The results are

		Huber/White method	
	Usual formula	Unclustered	Clustered
Average estimated variance of the coefficient			
var(_cons) $\times 10^4$	4.000	3.986	5.883
var(_b[x]) $\times 10^4$	11.99	11.93	17.53
correlation	0.	0.	0.
RMS error of the variance estimate			
var(_cons) $\times 10^8$	32	32.	4.
var(_b[x]) $\times 10^8$	96	96.	14.
correlation	0	.030	.064
Percent of cases marked as significant			
var(_cons)	10.3	10.3	5.0
var(_b[x])	10.5	10.7	5.3

The usual estimates did not change, and neither did the Huber results uncorrected for clustering. However, they are no longer correct. The Huber estimates with correction for clustering got the right answers. The impact on the Type I errors is remarkable.

It is noteworthy that the DEFF approach to this problem did not predict the relative loss of precision. Evidently, this problem—although seemingly simple—is too hard for DEFF. The Huber answers did a much better job.

Experiment 3

Now we will try a small-sample model. We will draw 25 observations in a manner similar to Experiment 1. The results favor the usual estimates.

		Huber/White method	
	Usual formula	Unclustered	Clustered
Average estimated variance of the coefficient			
var(_cons) $\times 10^4$	.040	.036	.023
var(_b[x]) $\times 10^4$	.111	.096	.051
correlation	0.	0.	0.
RMS error of the variance estimate			
var(_cons) $\times 10^4$	35	111.	245.
var(_b[x]) $\times 10^4$	96	396.	751.
correlation	0	.265	.522
Percent of cases marked as significant			
var(_cons)	5.0	5.7	19.6
var(_b[x])	5.0	6.4	26.3

The Huber standard errors are notably smaller then the usual ones. Once again, we see that the Huber covariance estimates are more variable, the clustered ones more than the unclustered. Since the Huber estimates are too low, they are too likely to declare a result significant.

The reason for this is that there are mathematical constraints on the residuals. Formally, they need to add to zero and be orthogonal to each $\mathbf{x}$. In the current problem, it is as if we were doing the regression on only 5 values; the sum of squared residuals would be only $3/5$ of the actual variance for those sums. In fact, this is what we observe for the intercept (.023 is about $3/5$ of .040).

A correction can be worked out as follows. For each observation, the variance of the residual is $\sigma^2(1-h_i)$, where $h_i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$. For two observations, the covariance of the residuals is $\sigma^2(-\mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j')$.

Thus, when the IID model holds, but a covariance matrix is estimated via Huber's method, the underestimation bias is

$$-(\mathbf{X}'\mathbf{X})^{-1}\Big(\sum_{j=1}^{C}\sum_{i=1}^{N_j}\sum_{m=1}^{N_j}\mathbf{x}_{ji}'\mathbf{x}_{ji}\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{jm}'\mathbf{x}_{jm}\Big)(\mathbf{X}'\mathbf{X})^{-1}$$

This formula cannot be computed by Stata, but a useful bound can be obtained by noting that the interior diagonal terms ($i=m$) are h_i; the off-diagonal terms will be less than $\sqrt{h_i h_m}$; and so a simple approximation to the result would be to add to each residual $\sigma\sqrt{h_i}$ before applying _huber.

I have modified hreg, creating hreg2 to calculate this quantity. A further useful observation is that we can bound the asymptotic convergence rate for the Huber estimates. That bound is

$$O\Big(\sum_{j=1}^{C}\frac{N_j^2}{N^2}\Big)$$

So, if no cluster is larger than 5% or so of the total sample, the standard errors will not be too far off because each term will be off by less than 1 in 400.

Experiment 4

Experiment 4 is a repeat of Experiment 3 except that I used the `hreg2` command included on the STB diskette:

		Huber/White method	
	Usual formula	Unclustered	Clustered
Average estimated variance of the coefficient			
var(_cons) $\times 10^4$	.040	.039	.039
var(_b[x]) $\times 10^4$	.111	.108	.108
correlation	0.	0.	0.
RMS error of the variance estimate			
var(_cons) $\times 10^4$	35	114.	207.
var(_b[x]) $\times 10^4$	96	436.	872.
correlation	0	.242	.310
Percent of cases marked as significant			
var(_cons)	5.0	4.7	7.6
var(_b[x])	5.0	5.2	11.4

Much better! This does a reasonable job of correcting the answer for this problem, but may be an overcorrection for variables where there is not a lot of intracluster correlation.

A further problem arises now that the variance is correct on average. In some sense we only have 5 observations—one for each cluster—so perhaps the *t*-statistic ought to have 5 degrees of freedom instead of 23. Recalculating the percent of cases in the clustered case using 5 would result in the last part of the table reading:

		Huber/White method	
	Usual formula	Unclustered	Clustered
Percent of cases marked as significant			
var(_cons)	5.0	4.7	3.4
var(_b[x])	5.0	5.2	6.3

This further adjustment works well in this case, bringing the Type I error probabilities back into line.

Conclusions

The formulas above imply that the bias exists in proportion to the square of cluster size in relation to sample size. As long as the largest cluster is 5 percent or less of the sample, this bias should be negligible.

In the case where the 5-percent rule does not hold, an adjustment is possible. On the STB diskette, I provide `hreg2.ado` as an alternative to `hreg` that makes this adjustment.

I have also shown a formula which in principle corrects this bias in all cases. However, this formula is not easily implemented at present.

References

Brook, R. H., J. E. Ware, W. H. Rogers, et al. 1983. Does free care improve adults' health? *New England Journal of Medicine* 309: 1426–1434.

Huber, P. J. 1967. The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1: 221–233.

Rogers, W. H. 1983. Analyzing complex survey data. Santa Monica, CA: Rand Corporation memorandum.

Rogers, W. H. and J. Hanley. 1982. Weibull regression and hazard estimation. *SAS Users Group International Proceedings.*

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–830.