Testing Homogeneity Of A Large Data Set By Bootstrapping

[1]**Morimune, K and** [2] **Hoshino, Y**

[1]Graduate School of Economics, Kyoto University
Yoshida Honcho Sakyo Kyoto 606-8501, Japan. E-Mail: morimune@econ.kyoto-u.ac.jp
[2]Graduate School of Economics, Kyoto University

Keywords: *Wu-Hausman test; Micro data; Bootstrapping; Sub-sample.*

1 EXTENDED ABSTRACT

It is not rare to analyze large data sets these days. Large data is usually of census type and is called the micro data in econometrics. The basic method of analysis is to estimate a single regression equation with common coefficients over the whole data. The same applies to other method of estimation such as the discrete choice models, Tobit models, and so on. Heterogeneity in the data is usually adjusted by the dummy variables. Dummy variables represent socioeconomic differences among individuals in the sample. Including the coefficients of dummy variables, only one equation is estimated for the whole large sample, and it is usually not preferred to divide the whole sample into sub-samples. Data is said to be homogenous in this paper if a single equation is fit to the whole data, and it explains socioeconomic properties of the data well. We may estimate an equation in each sub-population if the whole population is divided into known sub-populations. It is assumed that the coefficients are different from one sub-population to another in this case. Data is said to be heterogeneous in our paper. The analysis of variance is applied if sub-populations are known and sub-sample is collected from each sub-population.

In this paper, a test is proposed to find if the data is homogenous or not. Our test uses the full sample of size N and randomly chosen sub-samples of size n. They are randomly chosen since sub-populations are unknown. A regression equation with common coefficients over the whole sample such as

$$y_{ik} = x'_{ik}\beta_0 + u_{ik}$$

is assumed under the null hypothesis. A regression equation with variable coefficients

$$y_{ik} = x'_{ik}(\beta_0 + \frac{n}{N}\beta_k) + u_{ik}$$

is assumed under the alternative hypothesis. This alternative hypothesis states that the deviation from a common regression is small when the size n of randomly chosen sub-samples is small compared with N. This reflects our intuition that it is too restrictive to fit one regression equation with common coefficients to a large sample. It may be impossible to avoid specification errors in this estimation. However, specification errors may be negligible if a regression equation is fit to a small sub-sample.

For a given sub-sample of size n, the Wu-Hausman statistic

$$WH = (b_s - b_f)'(V(b_s) - V(b_f))^{-1}(b_s - b_f)$$

is used for the test where b_f and b_s are the full sample and the sub-sample least squares estimator, respectively. It is asymptotically distributed as $\chi^2(K)$ under the null hypothesis where K is the number of coefficients. The sub-sample of size n is repeatedly and randomly taken from the full sample of size N for Ns times, and the test statistic is calculated for Ns times accordingly. Since n is arbitrary, various values of n are chosen in the test starting from 5% to more than one third of the full sample. An alternative WH test statistic uses the bootstrapping estimators of the coefficients and the variance covariance matrices.

The sub-sample test statistics can be correlated with each other since the sub-samples are randomly chosen from the full sample and can be overlapped. Critical values of the test statistics are calculated by simulations. An example follows.

2 INTRODUCTION

The population Π is partitioned into m sub-populations such as

$$\Pi = \{\Pi_1 \cup \Pi_2 \cup \cdots \cup \Pi_m\}.$$

The sample S consists of N subjects, and is partitioned into m disjoint sub-samples such as

$$S = \{S_1 \cup S_2 \cup \cdots \cup S_m\}.$$

The researcher, however, does not have any information on the partitions of Π nor S. The full sample includes observations on

$$(y_i, x_i), i = 1, 2, \cdots, N$$

where N is the full sample size, and the kth sub-sample is

$$(y_{jk}, x_{jk}), j = 1, 2, \cdots, n_k.$$

A regression equation

$$y_i = x_i'\beta_0 + u_i \tag{1}$$

is maintained over the whole sample under the null hypothesis. The error term satisfies usual assumptions, and $V(u_i) = \sigma^2$.However, it seems too restrictive to assume that the coefficients are fixed over the whole sample, in particular, when the data set is large.

For each sub-population,

$$\begin{aligned} y_{ik} &= x_{ik}'(\beta_0 + \beta_k) + u_{ik}, \\ i &\in S_k, k = 1, 2, \cdots, m \end{aligned} \tag{2}$$

is a possible regression equation under the alternative hypothesis, additional coefficients β_k are nuisances, and x_{ik}' is $1 \times K$ row vector of explanatory variables associated with the kth sub-population. However, there is no way to estimate coefficients consistently since the partition of the population is unknown. A feasible regression equation under the alternative hypothesis may be

$$\begin{aligned} y_{ik} &= x_{ik}'(\beta_0 + \frac{n}{N}\beta_k) + u_{ik}, \\ i &\in S_k, k = 1, 2, \cdots, m \end{aligned} \tag{3}$$

where n is the sub-sample size which are randomly chosen in the test. This specification implies that the nuisance parameter depends on n proportional to N, and it is negligible when this ratio is small. Motivation of this study lies in this equation. Nuisance parameters may be negligible if a randomly chosen sub-sample is relatively small. It may not be negligible if it is applied to large samples such as census. We will propose a test for this conjecture.

3 PROPERTIES OF ESTIMATORS

Denote b the least squares estimator of the slope coefficient. If we estimate the regression equation using the full sample where n is N, the full sample estimator is inconsistent under the alternative hypothesis, $i.e.$,

$$p \lim_{N\to\infty} b_f = \beta_0 + \lim_{N\to\infty} \sum_{k=1}^{m} (X'X)^{-1} X_k' X_k \beta_k$$

where

$$X' = (X_1', X_2', \cdots, X_m'), X'X = \sum_{k=1}^{m} X_k' X_k,$$

and

$$\sum_{k=1}^{m} (X'X)^{-1} X_k' X_k = I.$$

The sub-sample is small relative to the full sample. It is further assumed that $n \to \infty$ as $N \to \infty$,and also

$$\lim_{N\to\infty} \frac{n}{N} = 0, \tag{4}$$

then the sub-sample estimator is consistent, $i.e.$,

$$\begin{aligned} p \lim_{N\to\infty} b_s &= \beta_0 + \lim_{N\to\infty} \frac{n}{N} \sum_{k=1}^{m} (X_s' X_s)^{-1} X_{sk}' X_{sk} \beta_k \\ &= \beta_0 \end{aligned}$$

where X_s is the explanatory variables in a sub-sample, $X_s' = (X_{s1}', X_{s2}', \cdots, X_{sm}')$, and X_{sk}' consists of sub-columns in X_k' associated with the kth group or zero if kth group is not in the sub-sample.

The regression equation (3) can be written as

$$y_{ik} = x_{ik}'\beta_0 + \frac{n}{N}\eta_{ik} + u_{ik}, \tag{5}$$

$$i \in S_k, k = 1, 2, \cdots, m \tag{6}$$

where η_{ik} is a idiosyncratic nuisance term $x_{ik}'\beta_k$. A more general interpretation can be given to the nuisance term in (5). Whatever the interpretation can be, the nuisance term is negligible in a small sub-sample where (4) holds, but not in the full sample. The probability limit of the least squares estimator is

$$\begin{aligned} p \lim_{N\to\infty} b_s &= \beta_0 \\ &+ \lim_{N\to\infty} \frac{n}{N} \sum_{k=1}^{m} (\frac{1}{n} X_s' X_s)^{-1} (\frac{1}{n} \sum_{i \in Sub} x_{ik}\eta_{ik}) \end{aligned}$$

where the last summation is over the sub-samples. The least squares estimator is consistent if the assumption (4) holds.

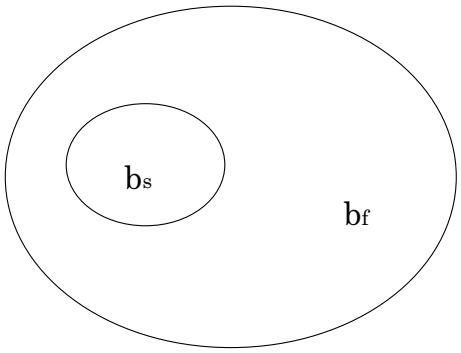
4 WU-HAUSMAN TEST

The null model of the test is equation (1), and the alternative model is equation (3) or (5). The Wu-Hausman test statistic is

$$WH = (b_s - b_f)'(V(b_s) - V(b_f))^{-1}(b_s - b_f) \quad (7)$$

where b_f and b_s are the full sample and the sub-sample least squares estimators, respectively. It is known that

Figure 1: Sub and Full Sample



b_f is the efficient estimator under the null hypotheses, and b_s is a consistent estimator under the alternative hypothesis. In fact, b_s is an instrumental variable estimator since

$$\begin{aligned} b_s &= (X_s'X_s)^{-1}X_s'y_s \\ &= (X'W(W'W)^{-1}W'X)^{-1}(X'W(W'W)^{-1}W'y) \end{aligned}$$

where the instruments are the selection matrix so that

$$W'X = X_s.$$

Then,

$$\begin{aligned} &V(b_s) - V(b_f) \\ &= \sigma^2\{(X_s'X_s)^{-1} - (X'X)^{-1}\} \end{aligned}$$

is positive definite. Moment condition is not satisfied by this instruments since $lim_{N\longrightarrow\infty}W'W/N = 0$.

The asymptotic distribution of WH is χ^2 with K degrees of freedom. This follows since, under the null hypothesis,

$$\begin{aligned} &Var\{\sqrt{n}(b_s - b_f)\} \\ &= \sigma^2\{(\frac{1}{n}X_s'X_s)^{-1} - \frac{n}{N}(\frac{1}{N}X'X)^{-1}\} \end{aligned}$$

and if the assumption (4) holds,

$$\begin{aligned} &\lim_{N\longrightarrow\infty} Var\{\sqrt{n}(b_s - b_f)\} \\ &= \sigma^2(\lim_{n\longrightarrow\infty}\frac{1}{n}X_s'X_s)^{-1} \end{aligned}$$

which is not degenerated. Note that $V(b_f)$ does not affect the asymptotic distribution, and $\lim_{N\longrightarrow\infty} Var\{\sqrt{N}(b_s - b_f)\}$ diverges to infinity. Furthermore, by the same reason,

$$\begin{aligned} \sqrt{n}(b_s - b_f) &= \sqrt{n}(b_s - \beta_0) - \sqrt{n}(b_f - \beta_0) \\ &= \sqrt{n}(b_s - \beta_0) - \frac{\sqrt{n}}{\sqrt{N}}\sqrt{N}(b_f - \beta_0) \\ &= \sqrt{n}(b_s - \beta_0) + o_p(1), \end{aligned}$$

then

$$WH = (b_s - \beta_0)'V(b_s)^{-1}(b_s - \beta_0) + o_p(1).$$

Under the alternative hypothesis,

$$\begin{aligned} p\lim_{N\to\infty}(b_s - b_f) &= p\lim_{N\to\infty}\{(b_s - \beta_0) - (b_f - \beta_0)\} \\ &= p\lim_{N\to\infty}\sum_{k=1}^{m}(X'X)^{-1}X_k'X_k\beta_k \end{aligned}$$

which is of O(1), and the consistency of the test is obvious.

Since this test depends on the selection of a sub-sample, sub-samples of the same size are chosen randomly and repeatedly for N_s times. These N_s test statistics are dependent on each other. For example, two test statistics are

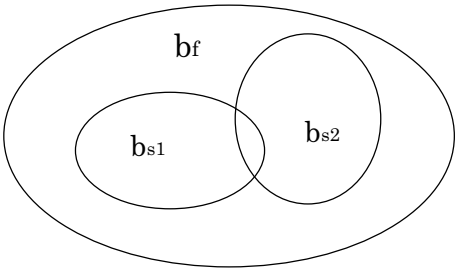
$$WH1 = (b_f - b_{s1})'(V(b_{s1}) - V(b_f))^{-1}(b_f - b_{s1}), \quad (8)$$

and

$$WH2 = (b_f - b_{s2})'(V(b_{s2}) - V(b_f))^{-1}(b_f - b_{s2}), \quad (9)$$

b_f is commonly used, and the two sub-samples S_1 and S_2 may share common observations.

Figure 2: Two Sub-samples



5 BOOTSTRAP TEST

The same hypotheses can be tested by the bootstrap method. Given a sub-sample, we estimate coefficient and the variance-covariance matrix by bootstrapping. Let B is the number of repetition in bootstrapping, then the coefficient is estimated by the sample mean

$$\bar{b}_s = \frac{1}{B}\sum_{i=1}^{B} b_{si},$$

and the variance covariance matrix is estimated by the sample moment

$$V(b_s) = \frac{1}{B-1}\sum_{i=1}^{B}(b_{si} - \bar{b}_s)(b_{si} - \bar{b}_s)',$$

and the full sample estimators $\bar{b}_f$ and $V(b_f)$ are calculated by the same way.

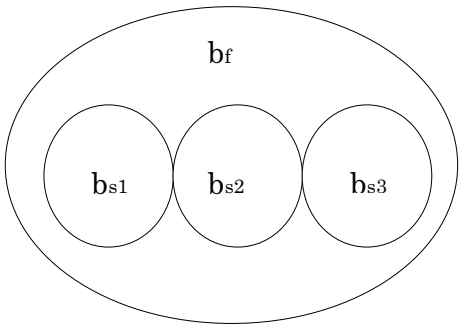
The hypothesis (1) can be tested for a particular sub-sample. However, if the test uses only one sub-sample, it will depend on a selected sub-sample and will be biased. To avoid the bias of the test, we randomly choose sub-samples from the full sample repeatedly for Ns times. The test is repeated for Ns times.

Since the test statistic is asymptotically distributed as χ^2 with K degrees of freedom, the empirical distribution of the test statistic is compared with the theoretical distribution. A simple method is to compare the real size of 5% test with the nominal size. If the empirical distribution rejects more than the nominal size, the null hypothesis of common coefficient is rejected.

6 BOOTSTRAP TEST OF INDEPENDENT SUB-SAMPLES

The bootstrap test explained so far uses dependent sub-samples. It is also possible to apply the same test statistic but using the non-overlapping sub-samples in the full sample. This method limits the number of sub-samples to N/n, and the computation is much faster. However, the computation can take long if the N/n sub-samples is repeatedly chosen. It is of interest to compare the dependent sub-sample and the independent sub-sample tests.

Figure 3: Independent Sub-samples



7 NULL DISTRIBUTION OF THE TEST

The null distribution is calculated for some cases by simulations. Dependency among the test statistics is of small order of magnitude $o((\frac{n}{N})^2)$, but it may affect the distributions of the test statistic. They depend on the following parameters.

1. Test statistics. (The WH test (7) or the WH test which uses bootstrapping estimations.)

2. Sample size N.

3. Sub-sample size n.

4. The number of coefficients K.

We have calculated the real size of 5% test under the null hypothesis of the test. The error term distribution is a normal distribution with unknown variances. The 5 percentiles of the χ^2 distribution with K degrees of freedom are used as the critical values. The table 1 tabulates real sizes of the WH test which uses the least square estimates of coefficients and variance covariance matrices. It may be found that the real size are very close to the nominal size.

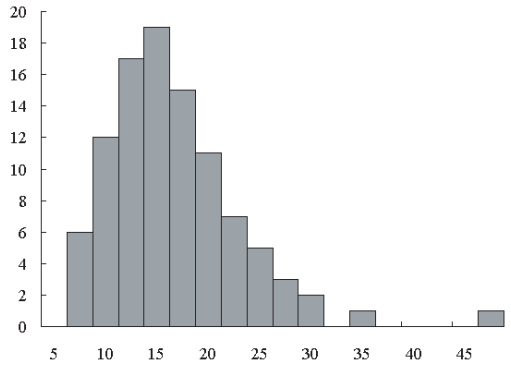
Table 1: Real Size of 5% WH Tests

n	K=5	K=10	K=15
1000	7	6	6
1500	5	6	6
2000	5	4	6
2500	5	5	6
3000	6	5	7
4000	6	5	3
5000	5	6	6
6000	5	4	7
7000	5	4	6
8000	6	5	5
10000	6	4	6

The upper 5% point of Chi-square (K) is used as a critical value.
(N=30000, Ns=800)

Histogram of WH test statistic is plotted for the case where N=30000, n=3000, K=15, B=200, and Ns=800. This empirical distribution is tested against the asymptotic null distribution $\chi^2(15)$. The null distribution was not rejected by Kolmogorov,Cramer-Von Mises, Watson, and Anderson-Darling tests. This result is natural since this simulation on the WH test statistic is almost the same as the simulation of χ^2 random variables. The difference is only in the estimation of variances of the error term.

Figure 4: Histogram of the WH test



In the table 2, real sizes of the bootstrapping WH test statistic are tabulated. Real size is mostly larger than 5%, and the dependency among the test statistics is not negligible when the sub-sample size is 8000 and 10000. The bootstrapping WH test has a thicker tail than the WH test statistic, and $\chi^2(16.8)$ as a null distribution cannot be rejected by the Cramer-Von Mises, Watson, and Anderson-Darling tests. This degrees of freedom is estimated by the method of maximum likelihood.

It may be more convenient to use the WH test than the bootstrapping WH test since the former does not need repeated calculations of bootstrapping. However, the null distribution of the WH test may heavily depend on the normal random variables.

Table 2: Real Size of 5% BWH Tests

n	K=5	K=10	K=15
1000	8	9	11
1500	7	7	11
2000	6	8	12
2500	7	8	11
3000	7	9	11
4000	7	9	10
5000	7	9	13
6000	7	9	14
7000	9	9	17
8000	8	11	18
10000	10	15	26

The upper 5% point of Chi-square (K) is used as a critical value.
(N=30000, Ns=800,B=200)

By examining the Tables 1 and 2, a proper sub-sample size in this test may be about ten percent or smaller of the full sample size. It will be found that it should not be too small since the bootstrapping method is degenerated.

8 EXAMPLE

We used the pair bootstrapping in our study since specification of the regression equation is in question in the test. Data is taken from Olsen (1998), and this example uses the probit estimation, not the linear regression, of a large sample. N=22272, K=19, and Nine independent variables are dummies. The ninth dummy is excluded in our study since it takes one only for 1241 individual among 22272, the ratio of which is 0.056. This dummy variable took only zero in sub-sample bootstrapping estimations a few times which terminated simulations.

The sub-sample size (n) is arbitrary. In this study, n is chosen to be from 1,000 to 10,000 which are from 5% to 50% of the full sample. Sub-samples, are randomly chosen. The number of sub-samples is 800 in the Wu-Hausman test. Since critical values of the WH test statistic are not calculated, the real size of the 5% χ^2 test are tabulated. (The second column in the Table 3. The critical value is 28.87 when the degrees of freedom is 18.) These real sizes are compared with the real sizes of the test statistics under the null hypothesis. (The third column in the Table 3. The critical value is 28.87, again.) Since the real sizes of the test statistics are mostly smaller than the real sizes of the null distribution, the null hypothesis may not be rejected. This data set is homogenous.

Table 3: 5% WH TEST

n	WH	null size
1,000	4.3	6
1,500	4.4	6
2,000	5.4	6
2,500	5.4	6
3,000	5.3	7
4,000	5.3	6
5,000	6.9	7
6,000	5.0	6
7,000	6.0	5
8,000	5.6	6
10,000	4.6	6

(N=22272, K=18 for each sub-sample. Ns is 800)

The bootstrapping WH test is also calculated and tabulated. The second column is the real sizes of the bootstrapping WH test statistics.(The critical value is 28.87.) The third column tabulates the real size of the test statistic under the null hypothesis. In this calculations, Ns and B are take to be 100 and 500, respectively, which turned out to be too small and too many, respectively. This means that the real sizes

of the test statistics are effected by the selection of 100 randomly chosen sub-samples, but they are stable even if the number of bootstrapping is reduced. On the whole, it seems this test does not reject the null hypothesis, either. The real sizes in the second column are smaller than those in the third column. It is noted that the second column shows dependency among test statistics as well as the third column when the sub-sample size is greater than 7000 which is one third of the full sample size. Compared with the Table 3, both the second and the third columns take larger values in the Table 4.

Table 4: 5% BWH TEST

n	real size	null size
1,000	3.8	11
1,500	4	11
2,000	8.5	12
2,500	8	11
3,000	9.4	11
4,000	8.6	10
5,000	7	13
6,000	8	14
7,000	11.3	17
8,000	13.5	18
10,000	22	26

(N=22272, K=18,B=500 for each sub-sample. Ns is 100*4)

The table 5 uses independent sub-samples. The first sub-sample is chosen randomly, and the second sub-sample is chosen randomly from the rest of the full sample. This continues until the rest of the full sample is smaller than n. Calculation is fast. In this calculation, the BWH test statistic takes more significant values than the table 4 shows, particularly when n is 2500 and 3000. It is necessary to repeat the calculation with different starting sub-samples, and the total task of calculation may not be anything faster than the dependent sub-sample test. The null distribution need to be calculated by the same way. However, the dependency among the test statistics will be avoided.

9 CONCLUSION

It was aimed to test the homogeneity of a census type large sample by the Wu-Hausman test statistic. This test statistic includes two sets of estimators as components. One is the full sample estimator usually used in empirical studies, and the other is a sub-sample estimator which uses only a part of the full sample. Naturally, the test is effected by the selection

Table 5: 5% TEST: Independent Sub−samples

n	Ratio of significant cases	significant cases	Ns
1,000	0.05	1	22
1,500	0.07	1	14
2,000	0.09	1	11
2,500	0.25	2	8
3,000	0.14	1	7
4,000	0.2	1	5
5,000	0	0	4
6,000	0	0	3
7,000	0	0	3
8,000	0	0	2
10,000	0	0	2

n:sub−sample size,Ns: Number of sub−samples. (N=22272,K=18,B=500)

of a sub-sample. We randomly choose the sub-samples, repeatedly calculate the test statistic, and examine the distribution of its values. It is necessary to study further properties of the test. 1. More precise null distributions are needed to derive the null percentile of the test statistic. 2. The power of the test must be examined. It is noted that the WH test is inconsistent under the usual specification of the alternative hypothesis (2). 3. The independent sub-sample and the overlapping sub-sample tests must be compared with each other from the view point of the null distribution and also of the power of the test. Most importantly, we need to develop a method to partition the full sample when the null hypothesis of a homogenous sample is rejected by the test. It seems too time consuming to measure distances among subjects in the sample.

10 REFERENCES

Olson, Craig A. (1998), Comparison of parametric and semiparametric estimates of the effect of spousal health insurance coverage on weekly hours worked by wives, *Journal of Applied Econometrics*, 13, 543-565.

Hausman, J. (1978), Specification Tests in Econometrics, *Econometrica*, 46, 1251-1271.