Journal of Statistical and Econometric Methods, Vol.12, No.1, 2023, 35-53
ISSN: 2241-0384 (print), 2241-0376 (online)
https://doi.org/10.47260/jsem/1213
Scientific Press International Limited

The Chow Test with Time Series-Cross Section Data

James K. Binkley[1] and Jeffrey S. Young[2]

Abstract

The Chow test is the standard method to test for differences in regression response across groups. In some cases, the groups being tested are composed of a time series of cross sections. For example, when testing for differences across industries, each industry may be composed of several observations on several individual firms. If the individuals themselves have systematic differences, the Chow test will be compromised: the individual and group effects become confounded. This can cause rejections in the absence of the group effect of interest. We illustrate the problem with a Monte Carlo analysis, and show that the effects cannot be separated. We propose a bootstrap-like testing procedure that can eliminate excessive Type I errors, and when used with the standard Chow test can help to arrive at an appropriate conclusion when both effects are present.

JEL classification numbers: C01, C10, C12, C15, C18.
Keywords: Chow test, Bootstrap, Empirical distribution, Statistical testing, Econometrics.

Researcher(s) own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business.
The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

[1] Professor Emeritus, Agricultural Economics, Purdue University.
[2] Assistant Professor, Agribusiness, Murray State University.

Article Info: *Received:* November 11, 2022. *Revised:* December 6, 2022.
Published online: December 12, 2022.

1. Introduction

Consider the problem of testing whether firms in the steel and chemical industries have different dividend policies. A common method for this is the well-known Chow test (1960), which tests for group effects by comparing the error sum of squares (ESS) from regressions on the individual industries to the ESS from a pooled regression using an F-test. It is usually characterized as involving two groups, but the test is easily extended to several groups. This makes it an attractive tool for testing group differences in a wide range of fields study from public policy to biomedical engineering research (e.g. Kartikasari & Merianti, 2016; Chen et al., 2019). Hence, any potential problems with or hidden violations of the test and its assumptions can have broad implications.

The interest in this study is using the Chow test for detecting group effects, and examining a relatively little-known issue with how this can be done. To this end, consider Chow's (1960) own example of comparing dividend behavior across the steel and chemical industries. Data for the test may consist of a single observation on many steel firms and many chemical firms. This is likely to require a large number of firms, which may be difficult to obtain, and perhaps impossible in some cases. In Chow's own example, there are only a limited number of steel/chemical firms in existence, at least firms large enough to have stock price trading data publicly available. The alternative is to have time series observations on m_1 steel firms and m_2 chemical firms, a time series of cross-sections. Data of this type is also likely with geographic data, when there is panel-type data on a set of states or countries and the interest is in determining whether there are regional differences.

It is this case, when there are multiple data points for each observational unit within a group, that is the focus of this paper. We show that, because the groups being examined (e.g. the chemical industry and the steel industry) are themselves composed of subgroups (chemical firms and steel firms), the results of the test are likely to be misinterpreted. In the extreme case, one may conclude there are group differences when in fact there are none: there are differences, but they have nothing to do with the hypothesis. The problem is that differences identified by the test may be due not only to the groups of interest but also to the subgroups. For example, there may be not only industry effects, but also firm effects. Finding a significant difference may indicate an industry difference, but it may also reflect idiosyncratic firm differences, of no interest to anyone except possibly the firms themselves. We demonstrate the nature of the problem with analytical arguments and a simple Monte Carlo analysis. We then argue that the two effects cannot be separately identified, making an accurate grouping test infeasible. We propose a solution to this problem using a bootstrap testing procedure, which is demonstrated with two applications using actual data. We end by presenting some results regarding power of the proposed procedure.

Our work is related to previous work concerning pooling data (Baltagi, Bresson, and Pirotte, 2008). However, that work has been more concerned with the consequences of improper pooling rather than the detection of differences.

Throughout we use the example of firms and industries, though obviously results apply to any case with groups composed of subgroups (henceforth, without loss of generality, "industries" and "firms", respectively).

1.1 Overview of the Chow Test

Consider a standard k-variable regression model $Y = \boldsymbol{X}\beta + e$, where e is the usual error term and k includes an intercept. Data is available from two distinct groups in the data, such as the industries in Chow's example, or two geographic regions. Denoting the groups as A and B, the interest is in whether the same equation applies to both. That is, testing the null hypothesis $\beta_A = \beta_B$, which is the hypothesis characterizing the Chow test. As suggested above, the standard procedure is to estimate three regressions: one with the A data, one with the B data, and one with the data pooled in a single regression. Then the Chow test statistic is as follows:

$$\frac{ESS_P - (ESS_A + ESS_B)}{ESS_A + ESS_B} \times \frac{n_A + n_B - 2k}{k} \tag{1}$$

where the ESS's are the error sum of squares from the regressions. The statistic has an F distribution with k and $n_A + n_B - 2k$ degrees of freedom. Note that, should the hypothesis be extended to finitely many, or m industries, then the last term being subtracted in the second degrees of freedom becomes $m * k$. In this example, $m = 2$ because there are two distinct groups (industries), making the subtracted term equal to $2k$.

An alternative method to conduct the test is to use intercept and slope shifters (Gujarati, 1970). To illustrate we use a simple univariate model $Y = \alpha + \boldsymbol{X}\beta$. Continuing the A-B distinction, we apply the test in the usual way. Thus, we estimate this equation three times, as described above. The alternative is to use the model $Y = \alpha + \boldsymbol{X}\beta + \alpha_1 D + \beta_1(D\boldsymbol{X})$, where D is a dummy variable indicating either of the industries, say, B. This is estimated once, using the pooled data. It yields an equivalent test because it yields coefficients numerically equivalent to those in the individual regressions. For example, $\hat{\beta}$ will equal the original $\hat{\beta}_A$; $\hat{\beta}_1$ will be the same as $\left(\hat{\beta}_B - \hat{\beta}_A\right)$, and similarly for the α's. It follows that an F test of $\alpha_1 = \beta_1 = 0$ in this model is a test that the coefficients for the industries are the same, that is, a Chow test.

For either method, the generalization to m industries is fairly obvious. One either estimates more individual regressions to obtain more individual ESS's, or one uses a larger set of intercept and slope shifters.

1.2 Testing with Time Series-Cross Section Data

We now consider the case of interest in this paper: testing when the cross sections being examined are each composed of time series. For convenience we continue the example of firms and industries. Suppose we have firm data on some relation, and it is of interest whether there are differences in response across industries, where each industry is a subset of firms. The vector of coefficients of firm i in industry g can be thought of as $\beta + \pi_g$, where β is the effect common to all firms in all industries and π_g is a k-vector of industry effects, specific to firms in industry g. The Chow test amounts to testing that $\pi_g = \pi_h = \cdots = \pi_p$, (or equivalently, they are all zero), where $g, h, \ldots, p$ account for all firms and each firm is in one of the industries. If the test is rejected, we conclude there are industry effects.

A problem arises if the coefficient vector for firm i in industry g is actually $\beta + \pi_g + \nu_i$, where ν_i is a k-vector of elements measuring systematic coefficient differences in firm i relative to the average firm, unrelated to industries. That is, ν_i does not vary in time within a firm. It seems reasonable that this would be a common occurrence: each member of a group would not be expected to respond *exactly* as do other members of the industry. In the example, if such individual firm effects are "large," rejection of the hypothesis may in part be due to some joint effect of the component firms, which could occur with any grouping of firms.

Consider the simplest possible model $y = \mu + e$, where μ is the unconditional mean and e is the usual error term. The OLS estimate is $\hat{\beta} = \bar{y}$. With industry effects, we have $\hat{\beta}_g + \bar{y} + \hat{\theta}_g$. If there are no firm effects, then $\hat{\theta}_g = \hat{\pi}_g$, the industry effect. With firm effects, $\hat{\theta}_g$ is an estimate of $\pi_g + \frac{1}{m}\sum_{i=1}^{m}(\nu_i)$, where m is the number of firms in g and we assume the sample size in each firm is the same. Since under this condition the Chow test is a test on the $\hat{\theta}$'s rather than the $\hat{\pi}$'s, if the second term on the right is important relative to the first, a deceptive outcome is likely. Obviously this is more likely when the ν's are "large" relative to π_g. We see from the definition of $\hat{\theta}$ it is more likely when n, the number of observations per firm is large, making firm effects more detectable. It may also be more likely when *m*, the number of firms in an industry, is small, since then positive firm effects are less likely to be counterbalanced by negative effects. However, this itself may be counterbalanced by a reduction in the total number of observations, making any differences less detectable. In any case, if there are no actual industry effects, the Chow test may still reject the hypothesis, due to firm effects in the specified industries. This occurs even though the effect has nothing to do with the underlying grouping criterion but will occur with randomly chosen firms, as we show later.

1.3 Identifying Industry Effects

In view of the foregoing, an important question is whether or not the group effects of interest and the underlying subgroup effects can be identified and separately tested in a traditional regression model, such as in a Least Squares Dummy Variable (LSDV) model. The answer to this question is "no". To illustrate, consider a very simple case. Suppose we again have the univariate model $Y = \alpha + \beta X$, and we have panel data on four firms, firms 1 and 2 from industry A and firms 3 and 4 from B. The interest is in whether there is an industry effect. Ignoring any firm differences, to test this we can employ the intercept-slope shifter method in the same vein as Gujarati (1970) and use the model $Y = \alpha + \beta X + \alpha_A D_A + \beta_A(D_A X_A)$, where D_A is a dummy variable indicating industry A. The test is $\alpha_A = \beta_A = 0$.
To allow for firm effects, it might seem we can use the model:

$$Y = \alpha + \beta X + \alpha_A D_A + \beta_A(D_A X_A) + \alpha_1 D_1 + \beta_1(D_1 X_1) + \alpha_3 D_3 + \beta_3(D_3 X_3) \quad (2)$$

where the subscripts are obvious. This model separates all firms, first by industry, then within each industry. However, under the null hypothesis of no industry effect, $\alpha_A = \beta_A = 0$, the model in (2) becomes $Y = \alpha + \beta X + \alpha_1 D_1 + \beta_1(D_1 X_1) + \alpha_3 D_3 + \beta_3(D_3 X_3)$, which does not fully separate the firms: firms 2 and 4 are now combined. This implies the hypothesis $\alpha_A = \beta_A = 0$ can also be interpreted as a test of equality between firms 2 and 4. The ambiguity in interpretation arises because in order to identify firms 2 and 4, α_A and β_A must be in the model. In fact, because the model in (2) with α_A and β_A does separate the firms, it is statistically equivalent to the model:

$$Y = \alpha + \beta X + \alpha_2 D_2 + \beta_2(D_2 X_2) + \alpha_1 D_1 + \beta_1(D_1 X_1) + \alpha_3 D_3 + \beta_3(D_3 X_3) \quad (3)$$

a model which explicitly differentiates firms, with no allowance for industries. Thus, the industry test $\alpha_A = \beta_A = 0$ in (2) is equivalent to the firm test $\alpha_2 = \beta_2 = 0$ in (3), which explicitly tests that firms 2 and 4 respond equally. Note that if firms 2 and 4 respond equally, then there must be no industry effect, given they are in different industries: although this may seem to be insufficient to conclude the absence of an industry effect, since firms 1 and 3 may still differ, and they are in different industries. But if 1, which is in the same industry as 2, differs from 3, which is the same industry as 4, and given that 2 and 4 do not differ, this difference in 1 and 3 must reflect firm differences. To eliminate this confounding, the model in (3) would need to also include the variables D_2 and $D_2 X$. But then we would have perfect multicollinearity: the parameters would be unidentified. We conclude that to identify industry effects while accounting for individual effects is not possible. They are confounded and hence not separable.

This would seem to eliminate the possibility of validly testing for industry differences when there are multiple observations on the sampled units and the units have systematic differences. Essentially, with such multiple observations, the standard test must be made under the maintained hypothesis of no individual effects. But of course, assuming does not make it so: a maintained hypothesis must have some credibility, and many cases of the kind considered herein do not. Thus, the solution requires a test which explicitly incorporates the possibility of individual effects into the maintained hypothesis. We propose such a test in the next section following a brief illustrative exercise.

2. Empirical Illustration

2.1 Monte Carlo

Here we further illustrate the issue with a simple Monte Carlo analysis. For convenience, we continue the industry example, focusing on a model with firm effects, but no industry effects. This makes the Chow test's performance easy to assess: the test should reject the null hypothesis with probability equal to a Type I error, or the size of the test. We employ the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \qquad (4)$$

Where x_1 and x_2 are uniform (0, 20) variables and e is a normally distributed error. A typical model coefficient is $\beta_i = \beta + \pi_{gi} + \nu_i$ where $\beta = 10$, $\pi_{gi} = 0$, and ν_i is normal (0, σ_v) with σ_v taking on different values; ν_i is a constant value for each coefficient in each experiment. For simplicity, we use two industries composed of a varying number of firms with a varying number of observations per firm. Since this exercise is meant only to be illustrative, we used only a small number of possible values of each of the four parameters of the experiment. They are listed in Table 1.

Table 1: Parameter values in the analysis

Parameter	Values
Observations per firm	10, 20, 30, 40, 50, 60
σ_v	1, 2, 3, 4, 5
Number of firms per industry	10, 20, 30
σ_{error}	60, 80

There are 180 different combinations possible. We conducted an analysis with each combination, performing 100 iterations in each case. For each iteration, a sample based on the particular set of parameters was generated.

For $\sigma_{error} = 60$ and $\sigma_v = 2$, the R^2 from OLS estimation of (4) is approximately 0.60. Also note that $\sigma_v = 2$ implies that approximately 90% of the time, ν_i lies within the interval [-2(1.67), 2(1.67)]. Then two industry regressions and a pooled regression were estimated, followed by the Chow test. Our measure of test performance is the percent of rejections.

The basic results obtained from each parameter value are presented in Table 2, which is actually a set of several one-dimensional tables – that is, we ignore how the effect of one parameter might depend on others.

Table 2: Percent of iterations resulting in rejection of null hypothesis (varying all 4 parameters)

Observations per firm	*10*	*20*	*30*	*40*	*50*	*60*
% of tests rejected	0.23	0.41	0.48	0.51	0.57	0.62
$\boldsymbol{\sigma_v}$	*1*	*2*	*3*	*4*	*5*	-
% of tests rejected	0.17	0.40	0.51	0.62	0.65	-
Number of firms	*10*	*20*	*30*	-	-	-
% of tests rejected	0.46	0.46	0.49	-	-	-
$\boldsymbol{\sigma_{error}}$	*60*	*80*	-	-	-	-
% of tests rejected	0.50	0.44	-	-	-	-

We found little evidence of interaction. Thus, the entries for a given parameter are the percentage of rejections for all experiments with the parameter set at the indicated values at $\alpha =$ 0.05. If the Chow test performed correctly, then all entries in the table should be approximately 0.05.

Briefly examining Table 2, we see considerable evidence that the test did not perform correctly: all entries exceed 0.05, in many cases by a large amount. The two most influential parameters are the observations per firm and the standard deviation of the firm effect. From the discussion above, both of these are what we expected. The error standard deviation is also influential, driving the number of rejections down as it increases. This is not surprising: the larger the error variance becomes, the less sensitive is any test. The number of firms per industry does not appear to be very influential, although we had no strong expectations about to the importance of its effect.

We can summarize the results by regressing the percent of rejections at $\alpha =$ 0.05 in each experiment on the parameters of each experiment. This yields:

$$\mathbf{1}\{Reject\ at\ \alpha = 0.05\} = \underset{(3.72)}{0.110} + \underset{(0.84)}{0.001} firms - \underset{(-7.78)}{0.030} \sigma_{error} + \underset{(30.60)}{0.007} obs + \underset{(42.17)}{0.111} \sigma_v \tag{5}$$

The t-statistics are listed in parenthesis, and the R^2 is about 0.17. This regression essentially mirrors what is in the table: the important factors are the relative size of firm effects and the number of observations per firm. The error standard deviation has less influence, and the number of firms per industry has virtually none. We also estimated this equation with interactions between all the variables. The coefficients were difficult to interpret (a result of multicollinearity) and the R^2 was not an improvement over (5) (0.1733 versus 0.1728). From this, we conclude that the effects essentially act independently of one another.

2.2 Descriptive Analysis

To further examine the importance of the number of observations we conducted an experiment with varying observations per firm but all other parameters fixed. There are two industries, each with five firms. The standard deviation of firm effects is 2 and the error standard deviation is 60. Figure 1 graphs rejections versus sample size, for three conventional significance levels.

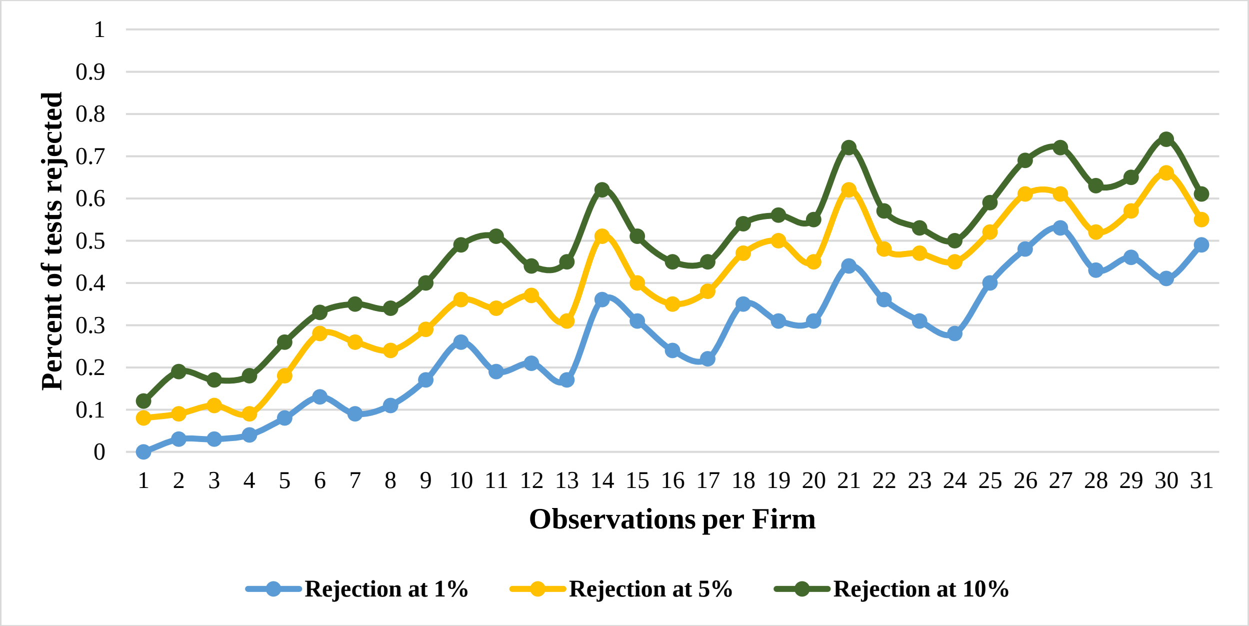

Figure 1: Rejections at the 1%, 5%, and 10% significance levels, by firm sample size

There is a clear response to increases in the number of observations. Each point is the average of 100 repetitions. As the number of observations per firm increases, the number of rejections also increases. From the graph, the slope appears to decrease as the number of observations increases. This is indeed the case; regressing the rejections on the observations in a quadratic model yields a significant quadratic term, and the model implies that at 60 observations per firm, there is no measurable effect brought about by additional observations. We also conducted a set of experiments comparing cases differing only by the number of firms, either five or ten, and found no discernible difference between the two. This supports the result that that the number of firms has little effect.

Thus, the evidence points to two key factors in determining the influence of firm effects in our problem: (a) the number of observations for each firm in the dataset, and (b) the magnitude of the firm effect. The number of firms has only a limited impact on the frequency of incorrect rejections in the Chow test. Hence, from a practical standpoint, it is better to have many firms, each with a small number of observations than to have a few firms, each with many observations.

As a final point, it is of some interest that as the error variance increases, that is, as R^2 declines, over-rejection from the Chow test becomes less of a problem. In our simulation, the R^2 was no larger than 0.60. In actual empirical studies, model fit is likely to be better than this. Therefore, we have if anything, understated the extent of over-rejection that is likely to occur in actual cases.

3. Methodology

3.1 Generating an Empirical F-Distribution

As just suggested, the problem can be framed as a failure of the assumption of no individual or firm-level effects. To eliminate this assumption, we propose a simple bootstrap-like procedure. To illustrate, we again use Chow's industry example. Suppose industry A is to be tested against industry B, and the data consists of t observations on each of n_A firms and n_B firms. The procedure is to generate a bootstrap F distribution, as follows. Step 1 is to create two artificial industries A' and B' by randomly assigning each of the $n = n_A + n_B$ firms to one of the two industries, requiring $n_{A'} = n_A$ and $n_{B'} = n_B$. Step 2 is to conduct a Chow test on these random industry groupings, generating an F statistic. Repeating this a large number of times generates a pseudo-F distribution. This distribution accounts for any firm heterogeneity that exists in the sample data, since it is based on the actual data generation process.

If there are firm effects, they will be in the random groupings to the same extent as in the correct industry grouping. The F statistic of interest, i.e., when the firms are correctly allocated to their industry, can then be compared to this empirical distribution. If there is no industry difference, this F statistic will be a random draw from the distribution. If there is an industry effect, the F statistic should be larger than a randomly drawn F from the empirical distribution. This follows since the correct allocation would be less heterogeneous: it not only accounts for any differences across the component firms, but also the industry differences. This procedure is termed "bootstrap-like" because resampling involves groupings of data (firms in the example), rather than individual observations, and the sampling is without replacement.

3.2 Power of the Procedure

These examples suggest that the procedure we have proposed is potentially useful in dealing with multiple-level effects when testing with the Chow test. We will now examine the power of the procedure, its ability to discern group effects when individual effects are present, and indeed when they are not. It is somewhat difficult to do this in any fully general way, due to the wide variety of model structures – the number of groups and individuals – under which the test may be applied. We will limit the analysis to three structures, each involving the standard case of two groups, which for convenience we again think of as industries, composed of firms. The three are 8 firms, with 4 firms in each industry; 20 firms, with 10 in each industry; and 20 firms, with 4 in one industry and 16 in the other. The regression model used is the same as employed in the Monte Carlo. As before, model coefficients can be expressed as $\beta + \pi_g + \nu_i$, where β is always 10; π_g is the group (industry) effect, taking values 0, 1, 2, 3, 4, and 5; and ν_i is a normal (0, σ_v) random variable representing a firm effect, with σ_v also taking integer values from 0 to 5. Experiments were conducted with 10, 20, and 30 observations per firm. Note that with $\beta = 10$, a value of, for instance, $\pi = 2$ implies a 20% increase in β. Also, $\pi = 2$ is "larger" than $v = 2$ in the sense that the average of the absolute value of a random variable distributed N(0,2) is less than 2.

The experiments proceeded as follows. For each of the three structures there are 108 combinations of parameters. For each of these, 50 samples were generated, and to each sample, the bootstrap procedure was applied. First a Chow test was performed with the firms correctly allocated to industries. Then the firms were randomly allocated to industries a number of times, each time conducting a Chow test. In experiments with 8 firms and two industries, the assignment was not actually random. This is because there are 35 unique ways to allocate 8 items to 2 groups. We simply conducted Chow tests for all 35, one of which was the correct allocation. For the experiments with 20 firms, the random assignment was repeated 60 times. Thus, each of the 50 runs generated either 61 or 35 F statistics, one of which was from the correct allocation of firms to industries. Call this F*. Our main interest is the location of F* in the empirical distribution of all the F's from that run, as in Figures 2 and 3. This is the basis of whether one accepts or rejects when using the bootstrap procedure.

The results for the three structures are presented in tables A1, A2, and A3 in the Appendix. In each case, we report (over the 50 experiments) the percent of times F* fell above the 75th and 95th percentiles of the empirical F distributions (percentages rounded to nearest integer). We also report the percent of F*'s that would lead to rejection ($p < 0.05$) by the standard Chow test. In order to avoid overdetailed tables, results for some values of π and v are not reported. A general point to be made here is that neither test performed well when the sample size is extremely small, bordering on inadequate for a meaningful estimation (e.g. $n = 10$). Although it is unlikely that many impactful studies would base their findings on samples as small as these, including such a case in the power simulations was necessary.

A brief examination of the tables shows that the results do not meaningfully differ across the three model structures. Thus, we will discuss them as a group. With no industry effect ($\pi = 0$) and with the presence of firm effects, the conventional test rejects too often, as much as 50 percent or more of the time when individual effects are strong, especially as sample size increases. This is what we observed in the Monte Carlo. It is not true of the bootstrap test. For example, the number of times F* falls in the upper quartile of the empirical F distribution does not often differ from the expected 25 percent, although there is considerable variability. However, the amount above the 95th percentile appears excessive in some cases, being as large as 16 percent. It is not clear whether this is anything beyond sampling variation. In any case, it is clear that use of the empirical F distribution is far less likely to result in mistaking firm effects for a non-existent industry effect.

Thus, the bootstrap method is effective in reducing Type I errors. The next question is how well it performs when the hypothesis is false. First consider the case where there is no individual firm effect, in which the textbook Chow test is fully applicable. In each table there are three examples of this, with industry effects of 1, 3, and 5, which together with 3 structures and 3 sample sizes generates 27 cases. Of these, the bootstrap procedure was worse than the conventional test (i.e. had fewer rejections) 10 times, better twice, and in the remainder they were equal. Power for both methods increased with n. With π taking values of 3 and 5, both tests rejected the hypothesis in nearly all cases.

With individual effects, performance of both tests declined. But the bootstrap test declined more. At high values of v, differences between the two tests were large in some cases. When π is very small relative to v ($\pi = 1$, $v = 5$), bootstrap performance was quite poor, with often little more than 25% of F* values falling above the 75th percentile. However, the standard test also had many failures in this case. Indeed, when the industry effect is dominated by firm effects in this manner, one might legitimately question whether there is a viable industry effect.

If π and v have the same value, both tests perform better the larger is the value. For example, performance is better when π and v are both 5 than when they are three. This occurs in all three tables, and suggests that the strength of the group effect is more important than the strength of the individual effect. Throughout, the bootstrap test is more likely to generate a Type II error, failing to detect a group effect.

The key difference between the two procedures is clearly their differing response to the value of v, the strength of individual effects within the groups being tested. The bootstrap is much more sensitive, and the effect is always negative. When there is no group effect, or when it is minor and economically inconsequential, this is desirable: it reduces Type I errors. But in the presence of both effects, it is overly conservative and hence prone to Type II errors. Thus, if only one test is to be relied on, which is better depends on the relative cost of Type I-Type II errors.

Of course, one need not rely on one test. The best way to proceed would seem to be to begin with a conventional Chow test. If it fails to reject the null hypothesis, there is no need for additional testing. If the hypothesis is rejected, and if it is reasonable to suspect there may be individual effects, then one is well advised to apply the bootstrap procedure and generate an empirical F distribution. With sufficient data for each unit, one can perform a standard Chow test within each industry to test for individual effects. If these are not rejected, one can conclude there are no individual effects and the original test is valid. If F* falls at a sufficiently high percentile (e.g., at least 0.9), this corroborates the conventional test. If F* falls well above the median but below conventional levels of significance, the combination of tests can reasonably be construed as indicating both industry and firm effects. An F* at or below the center of the distribution, regardless of how highly significant it may be, is best interpreted as evidence of strong firm effects, and no industry effect.

As a final point prior to application, it is well known that the Chow test is compromised if the individual regression models have different error variances. There is a large literature on the problem of heteroskedasticity in the Chow test, and alternative tests have been developed (Toyoda, 1974; Schmidt and Sickles 1977). Nevertheless, the problem is easily remedied by dividing each individual data set (e.g., for a firm) by the square root of the estimated error variance of the regression estimated with that data, that is, using weighted least squares. Note in this case the underlying heteroskedasticity model is known, eliminating the possibility of misspecification. This simple procedure has been found to be at least as good as more elaborate methods under most conditions (Thursby 1992). In all our tests involving actual data (discussed below), we performed this data weighting, without a prior test for heteroskedasticity.

4. Results

4.1 Application to Grocery Pricing Behavior

To illustrate, we now apply our procedure to two data sets. The first application of our procedure uses milk pricing data derived from the Nielsen Consumer Homescan Panel to examine a question addressed in the marketing literature: do chains in the same metropolitan market price similarly (Shankar & Bolton, 2004)? This is a natural candidate for the Chow test. Since most applications would not have data as extensive as ours, we used a randomly selected subset: 10 markets and 12 consecutive months (the same for each market). Of course, 10 markets is still quite large: most applications involve two. We tested the hypothesis with the stores correctly allocated to their markets, and a market fixed effect. The calculated F was 2.22 with 10 and 220 degrees of freedom, which is highly significant ($p < 0.001$), indicating similar pricing in local markets. We then generated an empirical F distribution by randomly allocating stores to the markets (with a market fixed effect), as described above. This is shown in Figure 2.

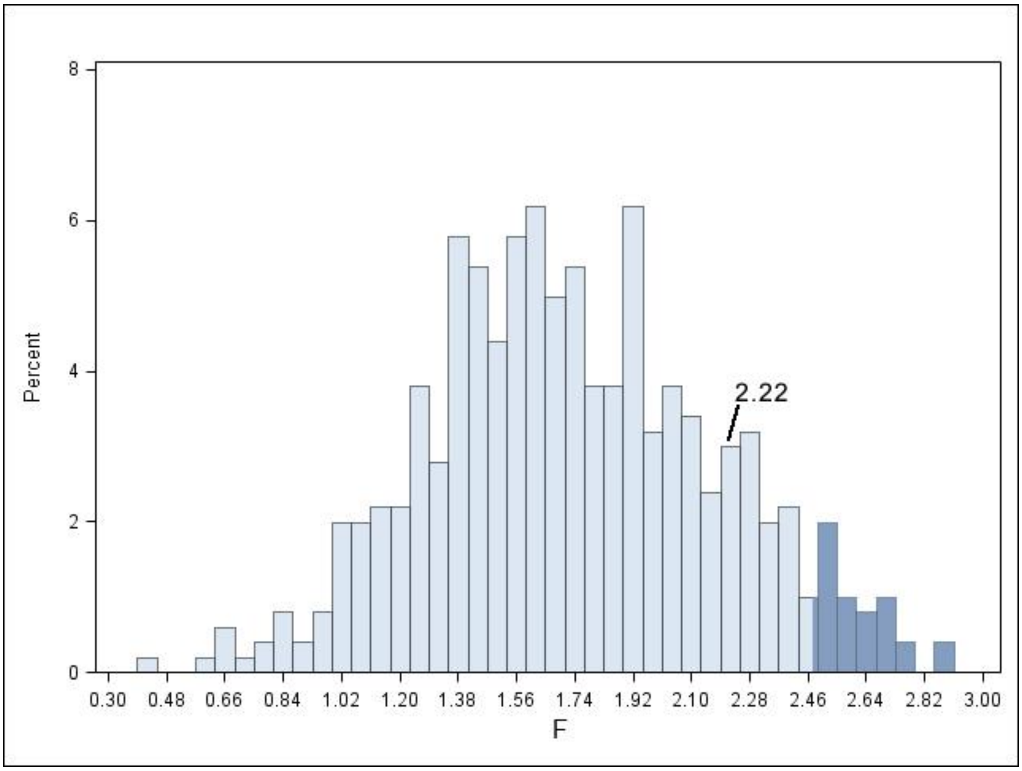

Figure 2: Bootstrap F distribution from regressions on the Nielsen data

The 95th percentile of this distribution is 2.50, with the darker area being the 0.05 rejection region; 2.22 is below this, around the 80th percentile. This would cast doubt on the original conclusion, and suggests firm effects in this sample. Indeed, 67 percent of the bootstrap F's were significant at 0.05 measured by the standard F. These results pertain to one set of 10 markets out of 49, over one of the 63 twelve-month periods. We repeated the procedure 25 times, each time randomly drawing a different set of 10 firms over a different 12 months, but with 100 iterations. In 20 of the 25 the hypothesis was rejected by a standard Chow test; but 5 of the 20 were not rejected by the bootstrap test. Thus in a large majority of these tests, the hypothesis would have been rejected by the standard test, but the fewer bootstrap rejections suggests that part of this strength is due to ignored firm effects. Indeed, 32 percent of the random groupings were significant at 0.05 using the theoretical F distribution, which shows the strength of these effects.

4.2 Application to Investment Activity

A possible problem with the method is that in many cases there may not be sufficient cross-section units (firms) to generate a usable bootstrap distribution. The range of combinations may be limited. For example, with two industries each with two firms, there are just three combinations. But a form of the method can still provide information about test validity. Hence, we turn to a second example, one similar to that given originally by Chow. The data is the Grunfeld investment data, which has been used in numerous econometric studies and is available in Greene (2012). The data involves eleven firms, each with 20 observations, with variables measuring investment, market value of stock, and real value of assets. The first is regressed on the other two. The data is useful here because among the eleven firms, there are four pairs for which each member of the pair is from the same industry. We used these eight firms to conduct a Chow test for an industry effect. Note this test involves four industries rather than the two in Chow's example.

First, we conducted the test with the firms correctly allocated to their industry. The calculated F was 33.98 with 3 and 148 degrees of freedom, which is highly significant by any standard when compared to the usual F distribution. This would seem to provide strong evidence of an industry effect. We then generated a bootstrap F distribution by randomly grouping the eight firms into four industries 400 times, repeating the Chow test each time, thus generating 400 F statistics. This empirical distribution is shown in Figure 3.

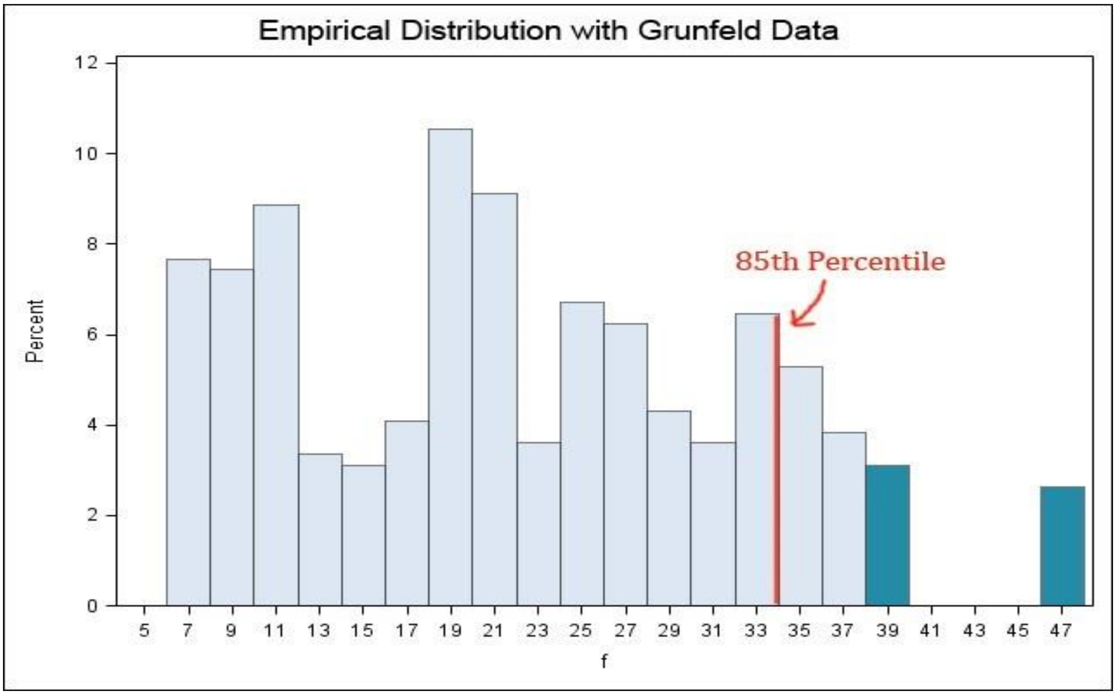

Figure 3: Bootstrap F distribution from regressions on the Grunfeld data

As indicated, 33.98 is in the 85th percentile of this distribution. This is not significant by normal standards. Thus, while there is some evidence of an industry effect, it is not strong, and not at the level suggested by the original F statistic. On the other hand, there is very strong evidence of a firm effect. Note from the figure that the smallest F statistic exceeds 6.0, which in this case has a probability value below 0.01. Some might argue that when conducting a Chow test, in many cases it makes more sense to permit intercept shifters for the industries and perform the test only on the slope coefficients (Wooldridge 2013). This amounts to applying the Chow test to fixed effects models. The bootstrap exercise was repeated for this case, and the F statistic for the correct pairing was in the 82nd percentile, providing yet weaker evidence of an industry effect.

One consideration in this case is that with only two firms and four industries, a large number of bootstrap samples is likely to have repetitions. That is assured here, for there are 115 unique possible pairings and we have 400 repetitions. However, repetitions do not invalidate bootstrap sampling. A more pertinent issue is that with only two firms per industries, many random pairings are likely to involve firms from the same industry, that is, correct groupings. With four industries, in any bootstrap sample we can have zero, one, two, or four such correct pairings. If samples with more correct pairings tend to have higher F values, this can be taken as evidence of an industry effect. We found some tendency for this in the samples. This is shown in Table 3, which has average F statistics for each of the four possibilities.

Table 3: F statistic descriptive data, by correct pairings

Correct Pairings	*n*		*F - Mean*	*F - Std. Dev.*	*F - Minimum*	*F - Maximum*
0	254		19.91	9.08	6.53	36.40
1	110		23.81	11.45	7.69	47.81
2	41		27.52	10.85	7.65	38.87
4	5		33.98	0.00	-	-

The average F consistently rises as the number of correct matches increase. But the table also shows that for this data, it is possible to get very large F’s even with no matches. In general, then, the table leads to the same conclusion as obtained above: weak but perhaps non-trivial evidence of industry effects and very strong evidence of firm effects.

A final point is that in this example, in which there are 115 possible groupings, the empirical F distribution has sufficient variability to enable meaningful probability statements. But if the number of unique pairings is very small, this may not be possible. But even with only a few different possibilities, meaningful conclusions may emerge. For example, with two firms and two industries, there are just three possible firm pairings: that with firms in the same industry grouped together, and the two incorrect pairings. We selected two industries from the Grunfeld data and conducted a Chow test for each of the three possible two-firm pairings of the four firms in these industries. When the firms were correctly paired by their industries, the F value was 5.98; the two incorrect pairings resulted in values of 2.10 and 6.59. All of these are “significant.” Since the F for the correct pairing falls between those for the incorrect pairings, neither of which can represent a test of industries, the appropriate conclusion would seem to be no real evidence of an industry effect, and (again) strong evidence of firm effects. This is similar to the previous case, and perhaps somewhat more definitive.

5. Conclusions

A question often encountered in econometric research is whether behavior differs across segments of a study population (in this paper, industries). To investigate this, one generally employs a Chow test, data for which consists of observations from the separate industries of interest. It may occur that these industries themselves are composed of multiple observations on component firms, such as several observations on each of several firms. In this paper we have shown that when this is the case, the Chow test can yield deceptive results (e.g. if region-level differences were of interest to a researcher, then sufficiently strong state-level heterogeneity could cause a Chow between states to overstate any differences, but it would be accounted for in our proposed testing procedure). This occurs when the firms themselves differ from each other, generating effects that become confounded with the effects of interest. As a result, the Chow test may reject the null when there are no differences in the industries of interest. In this paper we analytically examined the nature of the problem, and demonstrated its consequences using a simple Monte Carlo analysis. We proposed a bootstrapping procedure to deal with the problem. Using actual data, we demonstrated it can be quite useful in reducing the effects of this confounding, thus reducing the danger of test misinterpretation.

Our recommendation in performing a Chow test was not to rely on the results of only one test, but to begin with a conventional Chow test. Failure to reject the null hypothesis would imply that there is no issue, and thus is no need for running an additional test. However, if the null is rejected, there is reason to suspect the presence of any individual effects, then we recommend the researcher apply our bootstrap procedure in order to generate an empirical F distribution and re-evaluate.

We ended with an examination of test power. We found that when there is no confounding, the bootstrap procedure performs with accuracy similar to the standard Chow test. When both types of effect are present, the power of both tests declines, but the bootstrap test is considerably less likely to detect the industry effect of interest. Thus, under the condition stated, the proposed test has lower power, making more Type II errors. We suggest that in order to avoid errors it is advisable that both tests be used.

As a final point, it is evident that the data generating process underlying our analysis is similar to that underlying random coefficient and varying parameter models, for which various estimation methods have been proposed. One such method is to employ hierarchical modeling (Bryk & Raudenbush, 2002; Erkan et al., 2016; Meager, 2019), which estimates coefficient differences across the firms. However, this requires more information than that available for a Chow test, and goes beyond the simple purpose of determining whether there are differences in behavior. Our model is more akin to that underlying Swamy's random coefficient model (Swamy, 1970). This suggests the possibility of adapting Swamy's method to the Chow test. We performed some preliminary examination of this question by using Swamy estimation to obtain the error sums of squares for the Chow test. The result was a test with little power. This is a topic for future research.

References

[1] Baltagi, B.H., G. Bresson, and A. Pirotte (2008). “To Pool or Not to Pool.” The econometrics of panel data (pp. 517-546). Springer, Berlin, Heidelberg.

[2] Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (Vol. 1). Sage.

[3] Chen, Y., Ju, L. A., Zhou, F., Liao, J., Xue, L., Su, Q. P., ... & Zhu, C. (2019). An integrin α IIb β 3 intermediate affinity state mediates biomechanical platelet aggregation. Nature materials, 18(7), 760-769.

[4] Chow, Gregory. (1960). Tests of Equality between Sets of Coefficients in Two Linear Regressions. Econometrica, 28(3), 591-605.

[5] Erkan, A., Fainshmidt, S., & Judge, W. Q. (2016). Variance decomposition of the country, industry, firm, and firm-year effects on dividend policy. International Business Review, 25(6), 1309-1320.

[6] Gujarati, D. (1970). Use of Dummy Variables in Testing for Equality between Sets of Coefficients in Two Linear Regressions: A Note. The American Statistician, 24(1), 50-52.

[7] Greene, W. (2012). Econometric Analysis, Boston, Prentice-Hall.

[8] Kartikasari, D. and Merianti, M. (2016). The effect of leverage and firm size to profitability of public manufacturing companies in Indonesia. International Journal of Economics and Financial Issues, 6(2), 409-413.

[9] Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. American Economic Journal: Applied Economics, 11(1), 57-91.

[10] Schmidt, P, and R. Sickles. (1977). Some Further Evidence on the Use of the Chow Test under Heteroskedasticity. Econometrica, 45(5), 1293-1298.

[11] Swamy, P.A.V. (1970). Efficient Inference in a Random Coefficient Regression Model. Econometrica, 38(2), 311-323.

[12] Shankar, V. and R. Bolton. (2004). An Empirical Analysis of Retailer Pricing Strategy. Marketing Science, 23(1), 28-49.

[13] Thursby, J. (1992). A comparison of several exact and approximate tests for structural shift under heteroscedasticity. Journal of Econometrics, 53(1-3), 363-386.

[14] Toyoda, T. (1974). Use of the Chow Test under Heteroscedasticity. Econometrica, 42(3), 601-608.

[15] Wooldridge, J. (2013). Introductory Econometrics, Fifth Edition, South-Western, Mason O.

APPENDIX

Table A1: Two industries, each with four firms

		10 Observations			20 Observations			30 Observations		
π	*v*	*F*	*Empirical F*		*F*	*Empirical F*		*F*	*Empirical F*	
		P=.05	P=.25	P=.05	P=.05	P=.25	P=.05	P=.05	P=.25	P=.05
0	0	10	29	5	4	24	4	8	28	5
0	1	8	27	5	10	26	5	10	26	2
0	3	28	32	5	42	29	7	55	28	7
0	5	37	24	7	51	34	9	62	28	3
1	0	16	45	15	38	73	35	55	78	49
1	1	23	48	15	45	59	20	48	56	24
1	3	38	37	10	61	38	9	57	32	8
1	5	38	26	8	54	29	8	69	30	8
3	0	95	100	89	100	100	99	100	100	100
3	1	94	98	80	99	99	89	100	100	96
3	3	68	63	29	89	78	41	91	70	33
3	5	66	53	20	77	52	15	82	51	18
5	0	100	100	100	100	100	100	100	100	100
5	1	100	100	98	100	100	100	100	100	100
5	3	96	95	71	99	97	82	99	97	77
5	5	88	76	34	94	82	45	96	79	40

Table A2: Two industries, each with ten firms

		10 Observations			20 Observations			30 Observations		
π	*v*	*F*	*Empirical F*		*F*	*Empirical F*		*F*	*Empirical F*	
		P=.05	*P=.25*	*P=.05*	*P=.05*	*P=.25*	*P=.05*	*P=.05*	*P=.25*	*P=.05*
0	0	8	28	6	2	30	6	2	32	6
0	1	14	22	8	12	26	14	12	30	8
0	3	16	20	4	34	26	8	48	34	10
0	5	46	36	10	48	24	10	71	20	4
1	0	42	72	40	66	82	68	82	98	78
1	1	44	60	38	70	82	52	76	82	66
1	3	40	46	16	46	34	10	72	54	20
1	5	40	26	8	60	30	18	68	34	10
3	0	100	100	100	100	100	100	100	100	100
3	1	100	100	100	100	100	100	100	100	100
3	3	90	94	76	100	98	92	96	92	82
3	5	72	66	54	98	86	48	94	74	50
5	0	100	100	100	100	100	100	100	100	100
5	1	100	100	100	100	100	100	100	100	100
5	3	100	100	100	100	100	100	100	100	100
5	5	96	94	86	98	96	90	100	98	84

Table A3: Two industries – one with four firms, the other with sixteen

		10 Observations			20 Observations			30 Observations		
π	*v*	*F*	*Empirical F*		*F*	*Empirical F*		*F*	*Empirical F*	
		P=.05	*P=.25*	*P=.05*	*P=.05*	*P=.25*	*P=.05*	*P=.05*	*P=.25*	*P=.05*
0	0	8	25	8	0	28	6	7	22	8
0	1	6	29	6	10	22	8	14	24	10
0	3	15	17	4	45	30	11	50	36	16
0	5	30	20	2	48	14	2	47	16	4
1	0	18	53	25	42	68	32	64	92	62
1	1	25	47	16	39	57	39	53	67	27
1	3	42	48	12	42	26	12	68	44	12
1	5	29	27	6	66	38	10	58	40	15
3	0	98	100	96	100	100	100	100	100	100
3	1	96	98	92	100	100	100	100	100	100
3	3	84	88	67	92	88	72	100	96	80
3	5	74	64	28	82	72	40	84	65	29
5	0	100	100	100	100	100	100	100	100	100
5	1	100	100	100	100	100	100	100	100	100
5	3	100	100	92	100	100	100	100	100	96
5	5	92	86	51	98	92	73	100	87	79