Journal of Economic Perspectives—Volume 17, Number 3—Summer 2003—Pages 177–194

Retrospectives
Who Invented Instrumental Variable Regression?

James H. Stock and Francesco Trebbi

This feature addresses the history of economic words and ideas. The hope is to deepen the workaday dialogue of economists, while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please write to Joseph Persky, c/o *Journal of Economic Perspectives*, Department of Economics (M/C 144), University of Illinois at Chicago, 601 South Morgan Street, Room 2103, Chicago, Illinois 60607-7121.

Introduction

The earliest known solution to the identification problem in econometrics—the problem of identifying and estimating one or more coefficients of a system of simultaneous equations—appears in Appendix B of a book written by Philip G. Wright, *The Tariff on Animal and Vegetable Oils*, published in 1928. Its first 285 pages are a painfully detailed treatise on animal and vegetable oils, their production, uses, markets and tariffs. Then, out of the blue, comes Appendix B: a succinct and insightful explanation of why data on price and quantity alone are in general inadequate for estimating either supply or demand; two separate and correct derivations of the instrumental variables estimators of the supply and demand elasticities; and an empirical application to butter and flaxseed. The great breakthrough of Appendix B was showing that instrumental variables regression can be used to estimate the coefficient on an endogenous regressor, something ordinary least squares regression cannot do, which makes instrumental variables regression a central technique of modern micro- and macroeconometrics.

■ *James H. Stock is Professor of Economics, and Francesco Trebbi is a graduate student in the Department of Economics, Harvard University, Cambridge, Massachusetts. Stock is also Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts.*

Perhaps because Appendix B differs so from the rest of the book, its authorship has been in doubt. There is, in fact, a plausible alternative author: Philip Wright's eldest son, Sewall, who by 1928 was already an important genetic statistician. Indeed, the second of the two derivations of the instrumental variable estimator in Appendix B uses the method of "path coefficients," which Sewall had recently developed (S. Wright, 1921, 1923). Some histories, including Goldberger (1972), Crow (1978, 1994) and Manski (1988), attribute Appendix B to Sewall Wright. Morgan (1990) and Angrist and Krueger (2001) attribute authorship to Philip, but opine that Sewall probably deserves some intellectual credit. Others, including Christ (1994) and Stock and Watson (2003), state that authorship is in question, but do not take a stand.

So who wrote Appendix B, and, by inference, who solved the identification problem in econometrics? The simplest way to solve this puzzle would have been to ask Sewall, but apparently nobody did, and he died in 1988.

Lacking eyewitnesses, we investigate this mystery by other means: searching for traces of literary fingerprints hidden in Appendix B. The field of stylometrics—the statistical analysis of literary styles—postulates that subtle differences in style among authors can be used to attribute texts of ambiguous authorship. A classic stylometric study is Mosteller and Wallace's (1963) authorship analysis of the unsigned *Federalist Papers*. More recently, Foster (1996) used stylometrics to attribute the authorship of the political novel *Primary Colors* to Joseph Klein, a charge he denied until confronted by the *Washington Post* with editorial corrections in his handwriting.

Our detective work entails using stylometric data (numerical measures of word usage and grammatical constructions) to assess whether Appendix B is most likely by Philip or Sewall—or, potentially, by neither. Although stylometrics sounds exotic, its main methods are just versions of standard econometric tools. In fact, our stylometric investigation provides a simple (and, we hope, fun) illustration of some econometric methods for analyzing high-dimensional data sets, in which the number of explanatory variables exceeds the number of observations. As we shall see, this econometric sleuthing clearly points to the true author of Appendix B.

The History of Instrumental Variable Regression and Appendix B

The first known publication in English to describe the identification problem in an empirical context was a book review by Philip Wright of Henry Moore's *Economic Cycles: Their Law and Cause* (Moore, 1914; P. G. Wright, 1915). Philip explained why what Moore famously called a "new type" of demand curve—an upward-sloping demand curve for pig iron—could just be the *supply* curve, traced out by a shifting demand curve. Philip's treatment was very brief (less than one page) and followed what must have been a difficult discussion of autocorrelations and frequency domain methods. In any event, Philip's analysis seems to have been largely overlooked, even though it is cited in E. J. Working's (1927) influential

exposition of the identification problem.[1] Philip Wright (1929) later elaborated on his 1915 analysis in his review of Henry Schultz's (1928) *Statistical Laws of Demand and Supply with Special Application to Sugar.*

One definition of instrumental variable estimation is the use of additional "instrumental" variables, not contained in the equation of interest, to estimate the unknown parameters of that equation. Thus defined, instrumental variables estimation predated Appendix B. As discussed in detail by Goldberger (1972), Sewall Wright (1925) used instrumental variables to estimate the coefficients of a multiple equation model of corn and hog cycles. In that work, he derived the instrumental variables estimating equations (the equations among correlations from which the coefficients of interest could in turn be estimated) using the method of "path coefficients," which he had recently introduced in S. Wright (1921, 1923).[2] But the equations in his 1925 model are not simultaneous and all the regressors are exogenous, so ordinary least squares would have sufficed; there was no identification problem to solve, so instrumental variables estimation was unnecessary and appears to have been merely a computational expedient.[3] Moreover, in his 1925 paper, Sewall stated (footnote 7) that his method of path coefficients, as it then existed, could not handle systems of simultaneous equations.

The idea that instrumental variables estimation can be used to solve the identification problem—that is, can be used to estimate the coefficient on an endogenous variable—first appeared in Appendix B. Elaborating on P. G. Wright (1915) (but not citing Working, 1927), the author first presented the now-standard graphical demonstration of why movements in demand and supply can produce an arbitrary scatterplot of price-quantity points, which will trace out neither supply nor demand unless one of the curves is fixed; his key figure is reprinted as Exhibit 1. Then (pp. 311–312):

> In the absence of intimate knowledge of demand and supply conditions, statistical methods for imputing fixity to one of the curves while the other changes its position must be based on the introduction of additional factors. Such additional factors may be factors which (*A*) affect demand conditions

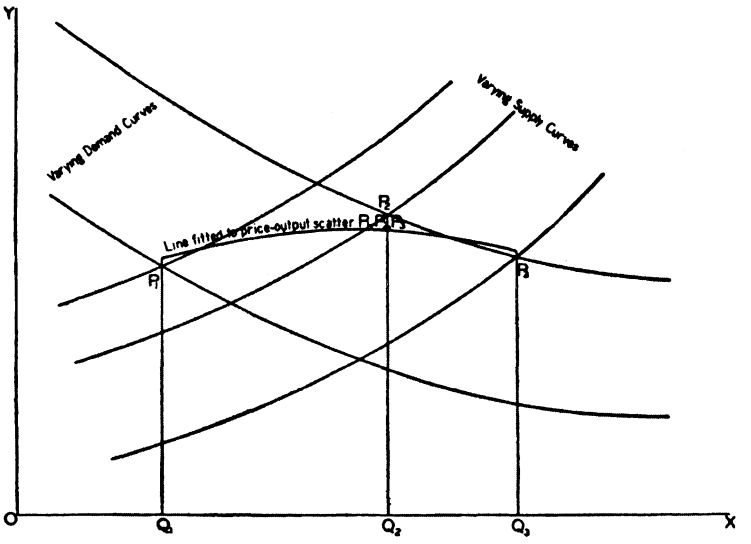
[1] Christ (1985, 1994) and Morgan (1990) provide engaging histories of the identification problem in econometrics and its solution. A single paragraph also suggesting that Moore had estimated a supply curve appeared later that year in Lehfeldt's (1915) review of Moore (1914). According to Christ (1985), the first known explanation of the identification problem was in French by Lenoir (1913) (translated as chapter 17 in Hendry and Morgan, 1995), but this is not referenced in other early work on this problem.

[2] The method of path coefficients begins by drawing a flow diagram with one-way arrows pointing from causal variables to intermediate variables to outcomes. This diagram allows one to trace the connection between any two variables by following the paths of arrows between them and produces a set of equations among correlations that can be solved to estimate the path coefficients. In Sewall Wright's (1921, 1923) initial expositions, the method of path coefficients is equivalent to multiple regression using ordinary least squares. Goldberger (1972) provides a clear discussion of path analysis and the estimation of path coefficients.

[3] Because S. Wright (1925) set to zero sample correlations that were nearly so, the instrumental variables estimating equations were simpler than the ordinary least squares equations in his four-regressor models.

Exhibit 1
The Graphical Demonstration of the Identification Problem in Appendix B (p. 296)

FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.



without affecting cost conditions or which (*B*) affect cost conditions without affecting demand conditions.

Appendix B then provides two derivations of the instrumental variable estimator as the solution to the identification problem. The first (pp. 313–314) is the "limited-information," or single-equation, approach, in which the instrumental variable *A* is used to estimate the supply elasticity; this derivation is summarized in Exhibit 2. The second derivation (pp. 315–316) is the "full-information," or system, derivation and uses Sewall Wright's (1921, 1923) method of path coefficients, extended to a system of two simultaneous equations. This derivation in effect solves the two simultaneous equations so that price and quantity are expressed as functions of *A* and *B*. Because *A* and *B* are exogenous, the resulting coefficients can be estimated by ordinary least squares, and thence, the supply and demand elasticities can be deduced. In modern terminology, this estimator of the elasticities is the indirect least squares estimator that, because the system is exactly identified, is the instrumental variables estimator obtained in the first derivation.[4]

The author of Appendix B refers to instrumental variable estimation as "the method of introducing external factors," which he then uses to estimate the supply and demand elasticities for butter and flaxseed. The external factors actually used

[4] From a modern perspective, the only flaw in the derivations is the loose treatment of the distinction between sample and population moments. This strikes us as a minor slip that can be excused by the early date at which Appendix B was written.

Exhibit 2
The Single-Equation Derivation of the Instrumental Variable Estimator in Appendix B

The derivation in Appendix B of the instrumental variable estimator of the coefficients of a single equation has two steps. The author tackled the supply curve first. Adopt his original notation, let O be the percentage deviation of output from its mean (then, as now, typically computed by taking the logarithm of the original quantity data, relative to its sample mean) and let P be the percentage deviation of price from its mean. Starting with the familiar supply and demand diagram, he first derived the supply curve with an additive disturbance,

$$O = eP + S_1,$$

where e is the elasticity of supply, S_1 represents the shift in the supply curve "brought about by a change in supply conditions," relative to when prices and output are at their long-run mean value, and the intercept is zero because the variables are deviated from their means. The author rearranges this expression as $eP = O - S_1$, then writes (p. 314):

> Now multiply each term in this equation by A (the corresponding deviation in the price of a substitute) and we shall have:
>
> $$eA \times P = A \times O - A \times S_1.$$
>
> Suppose this multiplication to be performed for every pair of price-output deviations and the results added, then:
>
> $$e \sum A \times P = \sum A \times O - \sum A \times S_1 \text{ or } e = \frac{\sum A \times O - \sum A \times S_1}{\sum A \times P}.$$
>
> But A was a factor which did not affect supply conditions; hence it is uncorrelated with S_1; hence $\Sigma\, A \times S_1 = 0$; and hence $e = (\Sigma\, A \times O)/(\Sigma\, A \times P)$.

(The shading has been added for the stylometric work carried out later.) The final expression for e is the formula for the instrumental variable estimator with a single instrument and a single included endogenous variable.

are not stated, but from context they appear to be the price of a substitute (A, which shifts demand) and the yield per acre (B, which shifts supply).

It is striking that Appendix B provides both limited-information (single equation) and full-information (system) derivations. Tinbergen (1930), apparently unaware of Appendix B, provided only one derivation, a full-information derivation

(using algebra, not path analysis) of the indirect least squares estimator.[5] The limited-information interpretation of the method of external factors apparently was not rediscovered, also independently, until the postwar work of the Cowles Commission.

Philip and Sewall Wright

Philip G. Wright (1861–1934) received a bachelor's degree from Tufts in 1884 and an M.A. in economics from Harvard in 1887.[6] Sewall Wright (1889–1988) was born in Massachusetts, and in 1892, the family (now including a brother, Quincy, born in 1890) moved to Galesburg, Illinois, where Philip became Professor of Mathematics and Economics at Lombard College, a small college that later folded in 1930. At Lombard, Philip taught economics, mathematics (including calculus), astronomy, fiscal history, writing, literature and physical education; he also ran the college printing press. Philip had a passion for poetry and used the press to publish the first books of poems by a particularly promising student of his, Carl Sandburg. Sewall graduated from high school in Galesburg in 1906 and attended Lombard College, where many of Sewall's courses, including his college mathematics courses, were taught by his father.

In 1912, Philip and Sewall moved to Massachusetts. Philip took a visiting position teaching at Williams College, and Sewall entered graduate school at Harvard. In 1913, Philip took a position at Harvard, first as an assistant to his former advisor, Professor Frank W. Taussig, then as an instructor. Taussig was subsequently appointed head of the U.S. Tariff Commission in Washington, D.C. In 1917, Philip left Harvard for a position at the Commission, then in 1922 took a research job at the Institute of Economics, part of what would shortly become the Brookings Institution. In 1915, Sewall received his Sc.D. from Harvard and took a position as Senior Animal Husbandman at the U.S. Department of Agriculture in Washington, D.C., where his responsibilities involved applying genetics to livestock breeding. In 1926, Sewall moved to the Department of Zoology at the University of Chicago, where he was promoted to professor in 1930.

When Philip had the time to write, he was prolific. While at Harvard, in addition to his 1915 review of Moore's book, he wrote a number of articles in the *Quarterly Journal of Economics*, and while at Brookings, he wrote several books and published articles and reviews in the *Journal of the American Statistical Association*, the *Journal of Political Economy* and the *American Economic Review*. Some of his writings

[5] Tinbergen (1930) discusses two estimators, the indirect least squares estimator and the "direct," or ordinary least squares, estimator. In his empirical application to the demand for potatoes, he *averages* the indirect least squares and ordinary least squares estimates of the demand elasticity. According to Morgan (1990, footnote 17, p. 182) and Magnus and Morgan (1987), at this point Tinbergen did not understand the statistical implications of the identification problem and saw no flaws with ordinary least squares estimation in simultaneous equations systems. Appendix B does not make this mistake.

[6] The biographical information in this section draws on Provine (1986), Crow (1994), Philip Wright's alumnus file archived at Harvard University and his personnel file archived at the Brookings Institution.

used algebra and calculus, typically following graphical expositions. Although Philip wrote on a wide range of topics, the identification problem was a recurrent theme in his work (P. G. Wright, 1915, 1929, 1930). In his later years, Philip was particularly concerned about tariffs, and he wrote passionately about the damage being done by recent tariff increases to international relations (P. G. Wright, 1933).

Sewall Wright became an eminent genetic statistician. In addition to developing the method of path analysis, he made fundamental contributions to evolutionary theory and population genetics. Evolutionary biology remained at the center of Sewall Wright's interests in his 76 years of publishing activity from 1912 to 1988, the year of his death at age 98. His only publications in economics were his 1925 analysis of the hog and corn markets undertaken as part of his duties at the USDA during World War I and a section of S. Wright (1934) that he coauthored with his father. According to Provine (1986, Table 1.2), Sewall expressed no more interest in economics than in Greek, Latin, astronomy or athletics.

Although Philip and Sewall may have experienced some tension over Sewall's choice of biology as a career (Provine, 1986), it appears that the two were intellectually close. In P. G. Wright (1915), Philip thanked Sewall for "valuable suggestions, and assistance in making the computations." Moreover, Philip and Sewall collaborated on a long section of a paper by Sewall explicating the method of path coefficients (S. Wright, 1934). That section elaborates upon the terse system derivation in Appendix B and shows that identification can be achieved by imposing other restrictions. In particular, they show that if there is only one instrument (for example, an instrument for supply, but not for demand), then system identification can be achieved by further assuming that the supply and demand errors are uncorrelated.

In short, it seems that either Philip or Sewall could have written Appendix B: Philip had a clear understanding of the identification problem as early as 1915, and Sewall's method of path coefficients was used in the second derivation of the instrumental variable estimator. If this contextual evidence does not resolve the mystery, perhaps textual evidence will.

Stylometric Analysis

When we started this project, we knew as much about grammar and style as most econometricians. Fortunately, we could draw on an established body of research that uses statistical methods to shed light on the authorship of disputed texts. The premise of stylometric analysis is that authors leave literary fingerprints on their work in the form of subconscious stylistic features that are largely independent of the subject matter. Father and son have many sole-authored publications, and they come from different generations and different literary traditions: Philip's passion was poetry; Sewall's, biology. Might quantifiable differences in their writing styles allow a clear attribution of Appendix B?

Stylometric analysis has three steps: collecting the raw texts, computing quantitative stylometric indicators and analyzing the resulting numerical data. In each, we break no new ground. For surveys of stylometric analysis, see Holmes (1998), Rudman (1998) and Peng and Hengartner (2002).

Data

The raw data consist of a sample of texts (listed in the Appendix) with sole authorship known to be Philip or Sewall, plus chapter 1 and Appendix B of *The Tariff on Animal and Vegetable Oils.* Photocopies of the originals were converted to text files using an optical character recognition program and checked for accuracy. The resulting text files were edited to eliminate footnotes, graphs and formulas. Following Mannion and Dixon (1997), blocks of 1,000 words were selected from these files. A total of 54 blocks were selected: 20 undisputedly written by Sewall, 25 by Philip, six from Appendix B and three from chapter 1. Although Philip's authorship of chapter 1 has never been questioned, we treat its three blocks as unknown to see if the authorship identification procedures correctly (we presume) attribute authorship to Philip.

We soon discovered that several of what we initially thought might be good stylometric indicators, such as sentence length and use of the passive versus active voice, have been found not to be useful for authorship identification because they are context specific or because they are subject to conscious manipulation by the author. Instead, the stylometric literature focuses on subtler elements of style (Rudman, 1994; Holmes, 1994, 1998). Rather than trying to develop our own stylometric indicators, we adopted two different sets of indicators directly from the literature.

The first set of stylometric indicators is the frequency of occurrence in each block of 70 function words. This list was taken wholesale from Mosteller and Wallace (1963, Table 2.5) and is presented in Table 1. These 70 function words produced 70 numerical variables, each of which is the count, per 1,000 words, of an individual function word in the block under analysis.[7] Because the word "things" occurred only once in the 45 blocks with known authorship, it was dropped from the data set, leaving 69 function word counts.

The second set of stylometric indicators, taken from Mannion and Dixon (1997), concerns grammatical constructions. Many of their indicators involved sequential word counts, for example, the average length of certain sentence segments. We decided that such length-based indicators could be unreliable in the context of mathematical writing (how many "words" is an equation?), so we did not compute them. Further excluding overlaps with the function words in Table 1 left 18 grammatical constructions; these indicators, which are either frequency counts per 1,000 or relative frequency counts, are listed in Table 2.

[7] Two blocks (items 3 and 7 under S. Wright's publications in the Appendix) were fewer than 1,000 words, so counts were scaled accordingly.

Table 1
Function Words Used in the Stylometric Analysis

a	all	also	an	and	any	are
as	at	be	been	but	by	can
do	down	even	every	for	from	had
has	have	her	his	if	in	into
is	it	its	may	more	must	my
no	not	now	of	on	one	only
or	our	shall	should	so	some	such
than	that	the	their	then	there	things[a]
this	to	up	upon	was	were	what
when	which	who	will	with	would	your

Notes: These are the function words listed in Mosteller and Wallace (1963, Table 2.5).
[a] Dropped from the data set because it occurred only once in the 45 blocks of known authorship.

Each 1,000-word block was processed to compute these stylometric indicators. The data set thus consists of 54 observations, each corresponding to a different block, on one dependent variable, authorship (Philip, Sewall or unknown) and 87 independent variables (69 function word counts and 18 grammatical statistics).[8]

Preliminary Data Analysis

Econometricians are trained to be skeptical. Is there any reason to think that these data can distinguish between Sewall's and Philip's known works, far less solve the mystery of Appendix B?

If a stylometric indicator differs substantially between the two bodies of known works, then it should be possible to detect that difference using a conventional differences-of-means *t*-statistic. As it happens, many of these *t*-statistics are large: of the 87 *t*-statistics (one for each indicator), 18 percent exceed 3 in absolute value and 41 percent exceed 2 in absolute value. So many large *t*-statistics would be quite unlikely if there truly were no stylistic differences between the authors and if the stylometric indicators were independently distributed.[9]

Table 3 presents summary statistics for the six stylometric indicators with the largest *t*-statistics; these indicators are the fourth grammatical statistic in Table 2 (a noun followed by a coordinating conjunction) and five function words. Evidently, Philip used the words "to" and "now" much more frequently than Sewall, while Sewall used the word "in" much more frequently than Philip.

Can we glean any preliminary indications of authorship from the counts in Table 3? One way to do so is to see whether the distribution of these indicators in

[8] Additional details on data collection and processing, the code in Perl used to compute the stylometric indicators, an electronic copy of Appendix B, a teaching note on classification analysis, and related material, are available by following the links from Stock's home page at ⟨http://post.economics.harvard.edu/faculty/stock/stock.html⟩.

[9] The indicators are not, however, independently distributed; for example, "now" and "then" tend to occur together, as do "if" and "would." Thus, formal joint inference on these *t*-statistics (such as a chi-squared test) is not straightforward.

Table 2
Grammatical Statistics Used in the Stylometric Analysis

occurrences of Saxon genitives forms 's or s'
noun followed by adverb
noun followed by auxiliary verb
noun followed by coordinating conjunction
coordinating conjunction followed by noun
coordinating conjunction followed by determiner
total occurrences of nouns and pronouns
total occurrences of main verbs
total occurrences of adjectives
total occurrences of adverbs
total occurrences of determiners and numerals
total occurrences of conjunctions and interrogatives
total occurrences of prepositions
dogmatic/tentative ratio: assertive elements versus concessive elements
relative occurrence of "to be" and "to find" to occurrences of main verbs.
relative occurrence of "the" followed by an adjective to occurrences of "the"
relative occurrence of "this" and "these" to occurrences of "that" and "those"
relative occurrence of "therefore" to occurrences of "thus"; 0 if no occurrences of "thus"

Notes: These grammatical statistics are the subset of those used by Mannion and Dixon (1997) after dropping statistics that overlap with Table 1 or are sequential word counts, which are ambiguous in mathematical texts.

Appendix B is closer to that in Philip's or Sewall's known texts. The results are suggestive: for all six indicators in Table 3, the mean and standard deviations of counts in Appendix B are quite similar to those found in Philip's writings, but different from those found in Sewall's.

Another way to get some insights into authorship is to see whether any of these "top six" indicators appear in the single-equation derivation quoted in Exhibit 2; as it happens, several do and are indicated by shading. The passage contains "now," a word used 1.6 times per 1,000 words by Philip, but only 0.1 times per 1,000 by Sewall. It also contains an instance, "deviations and," of a noun followed by a coordinating conjunction, a construction used almost twice as frequently by Philip as by Sewall, and it contains the word "to," which is used almost 50 percent more often by Philip than Sewall. On the other hand, the passage also contains the word "in," which is used more frequently by Sewall than by Philip. While this preliminary analysis points toward Philip as the author of Appendix B, it is not decisive. For firmer evidence, we must examine the full data set, but to do so we need different techniques.

Empirical Methods

An econometrician's first instinct might be to regress the binary authorship variable on the stylometric indicators. But with 87 regressors and only 45 observations, instinct soon gives way to reason: somehow, the number of regressors must be reduced before analyzing authorship. Two ways to handle this "dimension" problem are principal components analysis and linear discriminant analysis.

Principal components analysis entails reducing a large number of regressors to

Table 3
Summary Statistics for the Six Stylometric Indicators with the Largest *t*-Statistics

	Philip		*Sewall*			*Appendix B*	
	Mean	*Standard Deviation*	*Mean*	*Standard Deviation*	*t*	*Mean*	*Standard Deviation*
noun followed by coordinating conjunction	26.8	7.0	17.3	4.6	5.55	27.0	5.0
to	29.5	5.8	20.9	6.1	4.79	28.0	8.6
now	1.6	1.5	0.1	0.3	4.74	1.1	1.0
when	2.4	2.1	0.3	0.7	4.72	1.8	1.2
in	22.7	5.3	29.8	5.5	−4.34	18.5	5.8
so	2.1	1.6	0.7	0.8	3.82	2.0	1.7
n	25		20			6	

Notes: The entries in columns 2 and 3 are the mean and standard deviations of the counts per 1,000 words of the stylometric indicator in column 1 in the 25 blocks undisputedly written by Philip Wright. Columns 4 and 5 contain this information for the 20 blocks undisputedly written by Sewall Wright. The next column contains the two-sample *t*-statistic testing the hypothesis that the mean counts are the same for the two authors. The final two columns contain means and standard deviations for the 6 blocks from Appendix B. Shaded indicators occur in the excerpt in Exhibit 2.

a few weighted averages, or linear combinations, chosen to capture as much of the variation in the regressors as possible. The principal components approach begins by standardizing each variable, that is, by subtracting its sample mean and dividing by its sample standard deviation. The first principal component is the linear combination of the variables with the maximum variance, subject to the restriction that the squared weights sum to one. This procedure tends to give greater weight to regressors that are highly correlated. The second principal component is the linear combination of the regressors that has the second highest variance and is not correlated with the first principal component. The third, fourth and additional principal components are calculated in the same way.[10]

For our main analysis, we regressed the binary authorship variable on the first four principal components of the grammatical statistics, then repeated this for the function words. This produced a pair of predicted values for each observation, known or not.[11] Authorship of an unknown block is assigned depending on

[10] Specifically, let X denote the $n \times k$ matrix of n observations on the k standardized regressors. The first principal component of X is the linear combination of the regressors, $X\alpha_1$, that has the largest variance, where α_1 is a $k \times 1$ vector of weights normalized so that $\alpha_1'\alpha_1 = 1$. Because the sample variance of $X\alpha$ is $\alpha' X' X\alpha/(n-1)$, maximizing this sample variance subject to $\alpha'\alpha = 1$ is equivalent to maximizing $\alpha' X' X\alpha/\alpha'\alpha$, which is done when α is the eigenvector of $X' X$ corresponding to its largest eigenvalue. The second principal component is the linear combination formed using the second eigenvector of $X' X$, and so forth. For applications of principal components analysis in the stylometric literature, see Burrows (1987), Holmes and Forsyth (1995) and Peng and Hengartner (2002).

[11] This procedure can be applied generally to prediction or forecasting when the number of regressors is large, relative to the number of observations. For example, Stock and Watson (1999, 2002) and Forni et al. (2002) report promising results for macroeconomic forecasts based on the principal components of many predictors.

whether its pair of predicted values is closer to the means for Philip's or Sewall's known blocks, where distance is measured using the inverse covariance matrix of the pair of predicted values for the respective author. Several variations on this approach are explored as robustness checks.

Our second method, Fisher's linear discriminant analysis, was used by Mosteller and Wallace (1963) to analyze the *Federalist Papers*, although it is used infrequently in econometrics. Like principal components analysis, linear discriminant analysis constructs a linear combination of the stylometric indicators that can be used to distinguish between the two authors. Unlike principal components analysis, the linear discriminant analysis weights are computed using data on authorship. The weight (w_j) placed on a given variable (X_j) in Fisher's linear discriminant is the difference in the means for that variable between the known works of Philip and Sewall, divided by the sum of the variances of that variable for the known works of Philip and Sewall; that is,

$$Z = \sum_{j=1}^{k} w_j X_j, \text{ where } w_j = \frac{\bar{X}_{j;P} - \bar{X}_{j;S}}{s^2_{j;P} + s^2_{j;S}},$$

where $\bar{X}_{j;P}$ and $s^2_{j;P}$ are the sample mean and variance of variable j among works known to be written by Philip, $\bar{X}_{j;S}$ and $s^2_{j;S}$ are defined similarly for Sewall, and k is the number of stylometric indicators. When differences in the means are large, the weights will tend to be large—so that indicators that are quite different between the two authors receive greater weight than those that are similar. If the indicators are normally distributed with the same variances for both authors, then Fisher's linear discriminant is the optimal Bayes procedure for classifying authorship (Duda and Hart, 1973).

The linear discriminant was computed separately for the function words and grammatical statistics, respectively producing Z^{FW} and Z^{GS}. An unknown work is assigned to an author if the point (Z_i^{FW}, Z_i^{GS}) is closer to the average for Philip or for Sewall, where distance is measured using the inverse covariance matrix for the relevant author.

Cross-Validation Analysis

We begin the empirical analysis by testing these methods using what is known as "cross-validation" analysis. The idea of cross-validation is to drop an observation with a known value of the dependent variable (authorship) and to predict that value using the other observations; doing so repeatedly for all the observations provides an estimate of the prediction error rate. Performing this "leave-one-out" cross-validation analysis here entailed 45 repeated analyses; in each, 44 known texts are used to predict authorship of the remaining "unknown" text. This produced 45 authorship estimates that, because authorship of the "unknown" text is actually known, can be used to estimate the accuracy rate of our full sample analysis.

The resulting estimated accuracy rates are summarized in Table 4. Depending on author and statistical method, the estimated accuracy rate is 100 percent (that

Table 4
Cross-Validation Estimates of Accuracy Rates of Assigned Authorship

	Principal Components Regression		*Linear Discriminant Analysis*	
	Predicted Author:		*Predicted Author:*	
True Author:	*Sewall*	*Philip*	*Sewall*	*Philip*
Sewall	100%	0%	90%	10%
Philip	0%	100%	0%	100%

Notes: Based on leave-one-out cross-validation analysis of 45 1,000-word blocks of known authorship.

is, all texts are correctly identified) in three of four cases and 90 percent in the remaining case. The cross-validation estimates of 100 percent accuracy seem unrealistically high. Still, these results confirm that Philip and Sewall had different writing styles that are effectively distinguished by the stylometric indicators.

A Full-Sample Analysis

We now turn to our main statistical analysis, in which all 45 known texts are used to compute the principal components regression coefficients and the linear discriminant analysis weights.

Figure 1 is a scatterplot of the predicted values of the binary authorship variable from its regression on the first four principal components of the grammatical statistics (Y axis) and the first four principal components of the function words (X axis). (These principal components respectively explain 56 percent and 32 percent of the variance of the grammatical statistics and function words.) Figure 1 shows a clear separation between the works of known authorship. This is consistent with the high cross-validation accuracy rates and with the authors having measurably different writing styles.

The Figure 1 scatterplot also contains predicted values for the six blocks from Appendix B and the three blocks from chapter 1. All the blocks from Appendix B fall within the cluster of points associated with Philip's works, assigning authorship of Appendix B to Philip. All the blocks from chapter 1 also fall within Philip's cluster, correctly (we presume) assigning its authorship to Philip.

Figure 2 presents the comparable scatterplot of (Z_i^{FW}, Z_i^{GS}), the values of the linear discriminant for the grammatical statistics versus those of the function words. The conclusions are the same as from Figure 1: the values for Appendix B and chapter 1 fall squarely within the cluster of Philip's known texts.

Robustness Checks

We conducted several robustness checks. First, Mosteller and Wallace (1963) computed the linear discriminant using the differences of the medians instead of

Figure 1
Scatterplot of Predicted Values from Regression on First Four Principal Components: Grammatical Statistics versus Function Words

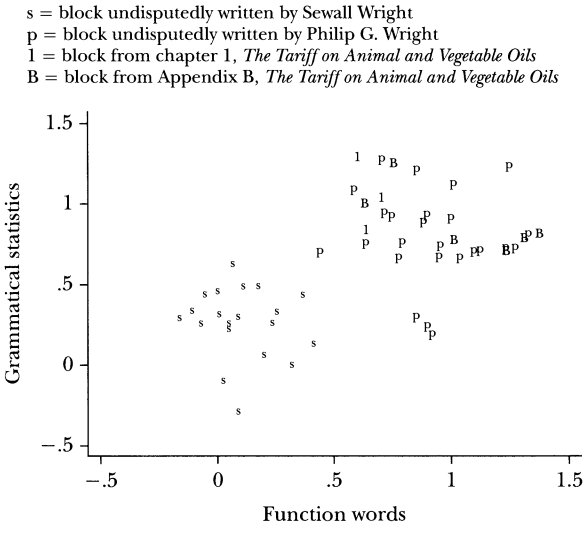
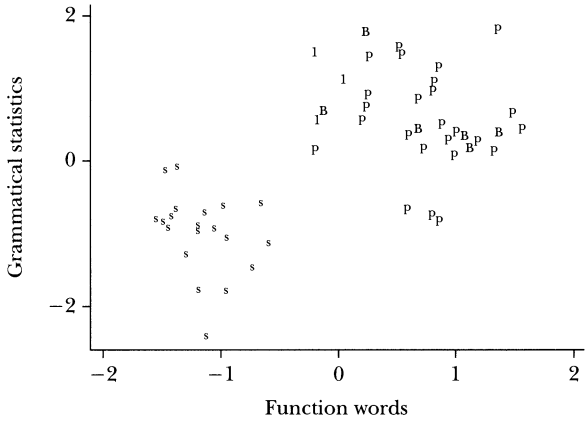


Figure 2
Scatterplot of Linear Discriminant Based on Grammatical Statistics versus Linear Discriminant Based on Function Words



the sample means and the squared ranges of the data instead of the sample variances, so we recalculated our linear discriminants using their alternative weighting scheme. The results are similar to those in Figure 2, assigning all the Appendix B and chapter 1 blocks to Philip.

Second, we computed the two principal components regressions using only the first two principal components, then again using the first six principal components. The results are similar to those in Figure 1, assigning all the Appendix B and chapter 1 blocks to Philip.

Third, we regressed authorship against an intercept, the first two principal components of the grammatical statistics and the first two principal components of the function word counts, and we attribute authorship depending on whether the predicted value is greater or less than 0.5. All works of known authorship were correctly assigned. All blocks from Appendix B and chapter 1 were assigned to Philip.

Fourth, we pooled all 87 stylometric indicators and computed their first four principal components (these explained 31 percent of the total variance) and assigned authorship first by regression, as in the preceding paragraph, and second by minimum distance in the resulting four-dimensional space. Again, all blocks of known authorship were correctly assigned, and all blocks from Appendix B and chapter 1 were assigned to Philip.[12]

Discussion and Conclusions

Who wrote Appendix B? The stylometric evidence clearly points to Philip G. Wright. Who first thought of using the instrumental variable estimator to solve the identification problem in econometrics? Of this we cannot be so sure: perhaps it was collaborative work or even Sewall's idea that Philip simply wrote up. Discussion of intellectual attribution, as opposed to authorship, quickly becomes speculative. Still, there is some relevant evidence.

In Sewall's favor, he was, after all, the inventor of the method of path analysis that was used in the second derivation of instrumental variable, and he had used an instrumental variables estimator in his earlier work on corn and hog cycles. Moreover, Sewall provided corrections to the first draft of Crow's (1978) biography of Sewall for the *International Encyclopedia of the Social Sciences*, in which manuscript Crow wrote of Philip Wright, "Later, in 1928, he wrote a book *The Tariff on Animal and Vegetable Oils*, to which Sewall contributed an appendix." While Sewall made other corrections to the entry, he did not amend this statement. As Crow pointed out in personal communication, however, Sewall missed a known factual error two sentences earlier, so perhaps (at an age of 88 years) his attention lapsed; alternatively, Sewall might have read "contributed an appendix" as "contributed to an appendix." Also, Arthur Goldberger brought a telling line to our attention: in a reprise, many years later, of the material in their 1934 coauthored section on supply and demand, Sewall Wright (1960, p. 431) wrote that "P.G. Wright [1928] . . . made, at my suggestion, a comparison of the results of this mode of approach [the method of path analysis] with results that he had arrived at by another method. . . ." Perhaps Sewall suggested the full-information derivation

[12] We also repeated the analysis using a different stylometric indicator developed by Benedetto, Caglioti and Loreto (2002), which uses zipped text compression ratios. This indicator also identifies Philip as the author of Appendix B. Their code is proprietary, and their published article is insufficiently detailed to permit replication, so we did not use this indicator for our main analysis.

using path analysis in Appendix B, even if Philip then carried out and wrote up the analysis.

In Philip's favor, it is evident from his 1915 book review that he clearly understood the identification problem and how it could be solved if one curve shifts while the other remains constant. Indeed, there are clear links between Appendix B and P. G. Wright (1915); in particular, Figure 3A in Appendix B is the same as Figure 3 in the 1915 book review, aside from unimportant differences in labeling. The first derivation of the instrumental variable estimator (the single-equation derivation) used graphical methods that would have been familiar to any economics instructor of the day, but we have not found any comparable derivations in Sewall's works. Although the full-information derivation used the method of path coefficients, it is plausible that Philip followed his son's research and saw its applicability to the identification problem. Also, as Crow pointed out to us, Sewall typically drew the published versions of path coefficients diagrams himself, but the path coefficient drawing in Appendix B (Figure 10) is not in Sewall's hand, rather, it was drawn by a professional draftsman. Finally, Sewall is not thanked anywhere in *The Tariff on Animal and Vegetable Oils.* Philip is not elsewhere chary with his acknowledgments: he thanks Sewall for suggestions and computational help in P. G. Wright (1915). At the time the book was written, Philip lived near Sewall and their families interacted (Provine, 1986, p. 102), yet Philip did not include his son among the dozen people he thanked in the acknowledgment section of the book.

In our view, this evidence points toward Philip as being both the author of Appendix B and the man who first solved the identification problem, first showed the role of "extra factors" in that solution and first derived an explicit formula for the instrumental variable estimator. Yet, as historians of econometrics like Christ (1985) and Morgan (1990) point out, a greater mystery remains: Why was the breakthrough in Appendix B ignored by the econometricians of the day, only to be reinvented two decades later?

Appendix
Analyzed Texts of Known Authorship

Sewall Wright

1. "Inbreeding and Homozygosis," *Proceedings of the National Academy of Sciences of the United States of America*, 19:4, pp. 411–20, April 15, 1933.
2. "Inbreeding and Recombination," *Proceedings of the National Academy of Sciences of the United States of America*, 19:4, pp. 420–33, April 15, 1933.
3. "Complementary Factors for Eye Color in Drosophila," in Shorter Articles and Discussion, *American Naturalist*, 66:704, pp. 282–83, May/June 1932.
4. "Statistical Methods in Biology," *Journal of the American Statistical Association*, Supplement: Proceedings of the American Statistical Association, 26:173, pp. 155–63, March 1931.
5. "Statistical Theory of Evolution," *Journal of the American Statistical Association*, Supplement: Proceedings of the American Statistical Association, 26:173, pp. 201–08. March 1931.
6. "The Evolution of Dominance," in Shorter Articles and Discussion, *American Naturalist*, 63:689, pp. 556–61, November/December 1929.

7. "The Dominance of Bar Over Infra-Bar in Drosophila," in Shorter Articles and Discussion, *American Naturalist*, 63:688, pp. 479–80, September/October 1929.

8. "Fisher's Theory of Dominance," in Shorter Articles and Discussion, *American Naturalist*, 63:686, pp. 274–79, May/June 1929.

9. "Effects of Age of Parents on Characteristics of the Guinea Pig," *American Naturalist*, 60:671, pp. 552–59, November/December 1926.

10. "A Frequency Curve Adapted to Variation in Percentage Occurrence," *Journal of the American Statistical Association*, 21:154, pp. 162–78, June 1926.

11. "Two New Color Factors of the Guinea Pig," *American Naturalist*, 57:648, pp. 42–51, January/February 1923.

Philip Wright

1. "The Bearing of Recent Tariff Legislation on International Relations," *American Economic Review*, 23:1, pp. 16–26, March 1933.

2. "Moore's *Synthetic Economic*," *Journal of Political Economy*, 38:3, pp. 328–44, June 1930.

3. "Cost of Production and Price," in Notes and Memoranda, *Quarterly Journal of Economics*, 33:3, pp. 560–67, May 1919.

4. "Value Theories Applied to the Sugar Industry," *Quarterly Journal of Economics*, 32:1, pp. 101–21, November 1917.

5. "Total Utility and Consumers' Surplus Under Varying Conditions of the Distribution of Income," *Quarterly Journal of Economics*, 31:2, pp. 307–18, February 1917.

6. "The Contest in Congress Between Organized Labor and Organized Business," *Quarterly Journal of Economics*, 29:2, pp. 235–61, February 1915.

■ *We thank James Crow for his recollections and for sharing his records with us, William Provine for checking his audiotaped interviews of Sewall Wright and Sarah Chilton of the Brookings Institution for archival research on Philip Wright. We are grateful to Carl Christ, Arthur Goldberger, Peter Reinhard Hansen, James Heckman and Mark Watson for helpful comments and discussions and to Vittorio Loreto for providing us with results from his zipping algorithm. We especially thank Alan Krueger for questioning, in a 2001 e-mail exchange, the first author's assumption that Sewall Wright wrote Appendix B; we decided solid evidence was needed. This research was supported in part by NSF Grant SBR-9730489.*

References

Angrist, Joshua D. and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives.* Fall, 15:4, pp. 69–85.

Benedetto, Dario, Emanuele Caglioti and Vittorio Loreto. 2002. "Language Trees and Zipping." *Physical Review Letters.* 88:4, pp. 048702-1–048702-4.

Burrows, John F. 1987. "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style." *Literary and Linguistic Computing.* 2:2, pp. 61–70.

Christ, Carl F. 1985. "Early Progress in Estimating Quantitative Economic Relationships in America." *American Economic Review.* December, 75, pp. 39–52.

Christ, Carl F. 1994. "The Cowles Commission's Contributions to Econometrics at Chicago, 1939–1955." *Journal of Economic Literature.* 32:1, pp. 30–59.

Crow, James F. 1978. "Wright, Sewall." Entry in the *International Encyclopedia of the Social Sciences–Biographical Supplement.* New York: Macmillan.

Crow, James F. 1994. "Sewall Wright," in *Biographical Memoirs of the National Academy of Sciences.* 64, pp. 439–69.

Duda, Richard O. and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis.* New York: Wiley.

Forni, Mario, Marc Hallin, Marco Lippi and Lucrezia Reichlin. 2002. "The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting." CEPR Discussion Paper 3432.

Foster, Donald. 1996. "Primary Culprit: An Analysis of a Novel of Politics." *New York Magazine.* February 26, 29:8, pp. 50–58.

Goldberger, Arthur S. 1972. "Structural Equation Methods in the Social Sciences." *Econometrica.* 40:6, pp. 979–1001.

Hendry, David F. and Mary S. Morgan. 1995. *The Foundations of Econometric Analysis.* Cambridge, U.K.: Cambridge University Press.

Holmes, David I. 1994. "Authorship Attribution." *Computers and the Humanities.* 28:2, pp. 87–106.

Holmes, David I. 1998. "The Evolution of Stylometry in Humanities Scholarship." *Literary and Linguistic Computing.* 13:3, pp. 111–17.

Holmes, David I. and Richard S. Forsyth. 1995. "The 'Federalist' Revisited: New Directions in Authorship Attribution." *Literary and Linguistic Computing.* 10:2, pp. 111–27.

Lehfeldt, Robert A. 1915. "Review of 'Economic Cycles: Their Law and Cause.'" *Economic Journal.* 25, pp. 409–11.

Lenoir, Marcel. 1913. *Etudes sur la Formation et le Mouvement des Prix.* Paris.

Magnus, Jan R. and Mary S. Morgan. 1987. "The *ET* Interview: Professor J. Tinbergen." *Econometric Theory.* 3:1, pp. 117–42.

Mannion, David and Peter Dixon. 1997. "Authorship Attribution: The Case of Oliver Goldsmith." *Statistician.* 46:1, pp. 1–18.

Manski, Charles F. 1988. *Analog Estimation Methods in Econometrics.* New York: Chapman and Hall.

Moore, Henry. 1914. *Economic Cycles: Their Law and Cause.* New York: Macmillan.

Morgan, Mary S. 1990. *The History of Econometric Ideas.* Cambridge, U.K.: Cambridge University Press.

Mosteller, Frederick and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association.* June, 58, pp. 275–309.

Peng, Roger D. and Nicolas W. Hengartner. 2002. "Quantitative Analysis of Literary Styles." *American Statistician.* 56:3, pp. 175–85.

Provine, William B. 1986. *Sewall Wright and Evolutionary Biology.* Chicago: University of Chicago Press.

Rudman, Joseph. 1994. "Nontraditional Authorship Attribution Studies in Eighteenth-Century Literature: Stylistics, Statistics and the Computer." Working paper.

Rudman, Joseph. 1998. "The State of Authorship Attribution Studies: Some Problems and Solutions." *Computers and the Humanities.* 31:4, pp. 351–65.

Schultz, Henry. 1928. *Statistical Laws of Demand and Supply with Special Application to Sugar.* Chicago: University of Chicago Press.

Stock, James H. and Mark W. Watson. 1999. "Forecasting Inflation." *Journal of Monetary Economics.* 44:2, pp. 293–335.

Stock, James H. and Mark W. Watson. 2002. "Macroeconomic Forecasting Using Diffusion Indexes." *Journal of Business and Economic Statistics.* 20:2, pp. 147–62.

Stock, James H. and Mark W. Watson. 2003. *Introduction to Econometrics.* Boston: Addison Wesley.

Tinbergen, Jan. 1930. "*Bestimmung und Deutung von Angebotskurven: Ein Beispiel.*" *Zietschrift fur Nationalokonomie.* 1, pp. 669–79.

Working, Elmer J. 1927. "What Do Statistical 'Demand Curves' Show?" *Quarterly Journal of Economics.* 41:1, pp. 212–35.

Wright, Philip G. 1915. "Moore's Economic Cycles." *Quarterly Journal of Economics.* 29:4, pp. 631–641.

Wright, Philip G. 1928. *The Tariff on Animal and Vegetable Oils.* New York: Macmillan.

Wright, Philip G. 1929. "Statistical Laws of Demand and Supply." *Journal of the American Statistical Association.* 24:166, pp. 207–15.

Wright, Philip G. 1930. "Moore's 'Synthetic Economics.'" *Journal of Political Economy.* 38:3, pp. 328–244.

Wright, Philip G. 1933. "The Bearing of Recent Tariff Legislation on International Relations." *American Economic Review.* 23:1, pp. 16–26.

Wright, Sewall. 1921. "Correlation and Causation." *Journal of Agricultural Research.* 20, pp. 557–85.

Wright, Sewall. 1923. "The Theory of Path Coefficients: A Reply to Niles' Criticism." *Genetics.* May, 8, pp. 239–55.

Wright, Sewall. 1925. "Corn and Hog Correlations." *U.S. Department of Agriculture Bulletin.* 1300, pp. 1–60.

Wright, Sewall. 1934. "The Method of Path Coefficients." *Annals of Mathematical Statistics.* 5:3, pp. 161–215.

Wright, Sewall. 1960. "The Treatment of Reciprocal Interaction, with or without Lag, in Path Analysis." *Biometrics.* 16:2, pp. 423–45.