

Testing the Statistical Significance of Linear Programming Estimators

Dan Horsky, Paul Nelson

William E. Simon Graduate School of Business Administration, University of Rochester, P.O. Box 270100,
Rochester, New York 14627 {horsky@simon.rochester.edu, nelson@simon.rochester.edu}

Linear programming–based estimation procedures are used in a variety of arenas. Two notable areas are multiattribute utility models (LINMAP) and production frontiers (data envelopment analysis (DEA)). Both LINMAP and DEA have theoretical and managerial advantages. For example, LINMAP treats ordinal-scaled preference data as such in uncovering individual-level attribute weights, while regression treats these preferences as interval scaled. DEA produces easy-to-understand efficiency measures, which allow for improved productivity benchmarking. However, acceptance of these techniques is hindered by the lack of statistical significance tests for their parameter estimates.

In this paper, we propose and evaluate such parameter significance tests. Two types of tests are forwarded. The first examines whether a model's fit is significantly reduced when an explanatory variable is deleted. The second is based on generating a standard deviation or distribution for the parameter estimate using nonparametric jackknife or bootstrap techniques. We demonstrate through simulations that both types of tests reliably identify both significant and insignificant parameters. The availability of these tests, especially the relatively simple and easy-to-use tests of the first type, should enhance the utilization of linear programming–based estimation.

Key words: attribute weights; DEA; linear programming; LINMAP

History: Accepted by Jagmohan S. Raju, marketing; received June 4, 2004. This paper was with the authors 9 months for 1 revision.

Introduction

Linear programming–based estimation techniques are used in a broad variety of settings, including marketing, operations, and accounting. However, not all the explanatory variables chosen a priori are likely to be relevant. Thus, despite some notable advantages over econometric analysis, wider acceptance of linear programming (LP) is hindered by the lack of statistical significance tests for its parameter estimates. Ideally, these tests should be simple and analogous to regression parameter t -tests. This paper addresses this stumbling block by developing and evaluating exactly such statistical significance tests for parameters estimated using LP in both the multiattribute utility function and multi-input production or cost frontier estimation settings.

Brand design, targeting, ad copy, and pricing decisions all require knowledge of the attribute importance weights consumers use to evaluate products. As a result, marketers commonly ask consumers to provide their preferences among a set of items to infer these weights. Regression often is used to perform this estimation because it is well known and has readily available parameter significance t -tests. With interval-scaled preference ratings, this is appropriate, but with ordinal-scaled data—a rank ordering

or paired preference comparisons—suspect parameter significance and fit statistics result. In addition, most consumers can confidently provide only ordinal-scaled preferences (Hauser and Shugan 1980). Hence, regression suffers from data quality problems if rating data are used and is inappropriate with ordinal preferences.

Alternatively, LINMAP (Srinivasan and Shocker 1973) utilizes only the ordinal properties of the preference data and performs well empirically relative to regression (e.g., Jain et al. 1979). In addition, Horsky and Rao (1984) show that LINMAP parameter estimates are consistent. However, LINMAP is not well known and does not provide parameter significance tests. Yet, given its favorable estimation qualities, an easily calculable test for parameter significance may lead to increased use. Even with this drawback, dozens of LINMAP-related applications exist (e.g., Kamakura and Srivastava 1986, Horsky et al. 2004). In addition, LINMAP is potentially applicable in a large array of fields such as psychology, medicine, the environment, and law that commonly utilize paired comparison data.

The analysis of production or cost as a function of input variables is commonplace in many literatures, in particular, economics, productivity, accounting,

and sales force. These literatures are dominated by extremal frontier applications of LP-based DEA that focus on the efficiency of particular decision-making units (DMUs). Central to DEA are the one-sided differences (i.e., the estimation errors) between the estimated frontier and actual output or cost levels. The efficiency of each particular DMU (plant or firm) corresponds to the size of this difference relative to the DMU's estimated frontier output or cost. Consequently, the efficient frontier serves as a "best-case" benchmark in performance evaluations.

DEA makes no distributional assumptions about the one-sided error terms. Thus, a standard-style statistical test for whether an estimated input parameter is statistically different from zero is unavailable. This drawback is a key argument used by Evans and Heckman (1988) in their objection to the use of LP techniques by Charnes et al. (1988) to analyze the breakup of the Bell System. Furthermore, because frontier estimation relies on extremal points, its results are very sensitive to variable selection (Seiford and Thrall 1990). Therefore the need for a statistical test of parameter significance (variable relevance) is magnified.

This paper develops and evaluates statistical significance tests for parameters estimated using LP in two contexts: (1) multiattribute utility function and (2) multi-input production or cost frontier estimation. We begin by reviewing the LP procedures used in both areas. Next, we develop statistical significance tests for the parameter estimates. Simulations then examine the ability of these tests to identify significant variables as significant (Type I error) and insignificant variables as insignificant (Type II error). We conclude with a summary and recommendations.

LP Estimation Procedures

The Multiattribute Utility Function

A brand's utility to a consumer depends on the brand's performance on a set of attributes. Either the linear or part-worth model typically is used to reflect this relationship. The true utility of brand i using the linear model is $\tilde{U}_i = \sum_{j=1}^J w_j b_{ij}$, where $j = 1, 2, \dots, J$ is the attribute index, w_j is the importance weight associated with attribute j and b_{ij} is the amount of attribute j contained in brand i .¹ Because of human

¹ Price is usually treated as an attribute. In the part-worth model, $\tilde{U}_i = \sum_{j=1}^J \sum_{k=1}^{K_j} \gamma_{jk} z_{ijk}$, where each attribute j has K_j possible levels and $z_{ijk} = 1$ if brand i has level k of attribute j , and 0 otherwise. γ_{jk} is the part-worth utility associated with level k of attribute j . The LINMAP formulation used to estimate the part-worth model is straightforward as are modifications of the statistical tests developed in the next section to evaluate the significance of the part-worth utilities.

error, the consumer's "stated" utility for item i differs from its true utility. That is, $U_i = \tilde{U}_i + \varepsilon_i$.

Estimation of an individual's utility function involves estimation of the attribute weights $\underline{w} = (w_1, w_2, \dots, w_J)$ based on the consumer's stated brand preferences (a proxy for U_i) and the attribute-level data (b_{ij}). LINMAP relies on the ordinal-scaled notion that if brand i is said to be preferred to brand k ; then the stated utility of i must exceed that of k . This inequality $U_i - U_k \geq 0$ becomes $\sum_{j=1}^J w_j (b_{ij} - b_{kj}) + \varepsilon_{ik} \geq 0$, where $\varepsilon_{ik} = \varepsilon_i - \varepsilon_k$. With N brands, there are $M = N(N - 1)/2$ paired comparisons each with an analogous inequality. Given these paired comparisons and the brands' attribute levels b_{ij} , LINMAP solves this system of linear inequalities for \hat{w} such that the estimated paired comparisons are as close as possible to the stated comparisons. Closeness is measured by the sum of the estimated brand utility differences $|\hat{U}_i - \hat{U}_k| = |\sum_{j=1}^J \hat{w}_j b_{ij} - \sum_{j=1}^J \hat{w}_j b_{kj}| = e_{ik}$ over the incorrectly estimated paired comparisons. Specifically, the LP is

$$\begin{aligned} \min_{\hat{w}} \quad & \sum_{i=1}^N \sum_{i \text{ preferred to } k} e_{ik} & (1) \\ \text{s.t.} \quad & \sum_{j=1}^J \hat{w}_j (b_{ij} - b_{kj}) + e_{ik} \geq 0 \\ & i = 1, 2, \dots, N, \text{ and all } k \text{ for which } i \text{ is} \\ & \text{preferred to } k, \\ & e_{ik} \geq 0 \quad i = 1, 2, \dots, N, \text{ and all } k \text{ for which } i \text{ is} \\ & \text{preferred to } k, \\ & \sum_{j=1}^J \hat{w}_j = 1. \end{aligned}$$

In effect, the e_{ik} are one-sided distribution-free error terms that are positive only if the paired preference comparison between brands i and k is incorrectly estimated, and equal zero otherwise.

The Production Frontier

Typically, researchers are interested in the estimation of an extremal relationship (efficient frontier) where maximum output is defined as a function of inputs. For simplicity, we assume that this production frontier is linear with respect to the inputs.² The true output possible by DMU i given the amounts c_{ij} of the inputs j it uses is thus $\bar{Q}_i = \sum_{j=1}^J v_j c_{ij}$, where v_j is the output associated with a unit of input j . Hereafter, we refer to the v_j as production coefficients.

² The cost function's development is analogous to that which follows and the same statistical tests developed in the next section apply. A more general nonparametric formulation of \bar{Q}_i is analogous to the part-worth utility model.

Since all or most DMUs are not completely efficient, actual production $Q_i = \hat{Q}_i - \varepsilon_i$, where $\varepsilon_i \geq 0$. If no assumptions are made about the distribution of these one-sided deviations from the production frontier, statistical estimation techniques are not applicable. Rather, LP-based DEA is used (Seiford and Thrall 1990). For the linear production frontier, the LP is

$$\begin{aligned} \min_{\hat{\theta}} \quad & \sum_{i=1}^N e_i & (2) \\ \text{s.t.} \quad & Q_i = \sum_{j=1}^J \hat{\theta}_j c_{ij} - e_i \quad i = 1, 2, \dots, N, \\ & e_i \geq 0 \quad i = 1, 2, \dots, N, \\ & \hat{\theta}_j \geq 0 \quad j = 1, 2, \dots, J. \end{aligned}$$

N is the number of DMUs. The e_i measure how far below the estimated efficient frontier firm i operates. As such this error directly measures the inefficiency of firm i , relative to its estimated maximal output, $\hat{Q}_i = \sum_{j=1}^J \hat{\theta}_j c_{ij}$. In fact, $\theta_i = e_i / \hat{Q}_i$ is referred to as firm i 's efficiency ratio.³

Statistical Significance Tests

Any estimation procedure results in parameter estimates. Statistical methods make a distributional assumption concerning the error terms. Thus the distributions of the parameter estimates are known asymptotically, the corresponding standard deviations can be calculated and a statistical significance test (e.g., a t -test) can be conducted. On the other hand, LP makes no distributional assumptions about the error terms. Hence, significance tests for the parameters obtained through the solution of problems (1) and (2) are not so straightforward.

We propose two types of tests to assess LP parameter significance. The first builds on the work of Goldfeld and Quandt (1972). They test the significance of nonlinear regression parameters using a likelihood ratio test to assess the reduction in fit caused by assuming that a particular parameter equals zero. It follows that the significance of an attribute weight or production coefficient can be tested by estimating problem (1) or (2) with attribute or input j and also without it (i.e., with $J - 1$ variables), and then examining the reduction in fit. In a similar way, hypotheses involving more than one parameter can

³ DEA commonly estimates the efficiency of each DMU separately. The envelope of the resulting input-maximum output relationships represents the estimated production frontier. This procedure allows even the most general nonparametric DEA formulations to be estimated using LP. The LP for the linear production function expressed by problem (2) corresponds to the output-oriented CCR ratio model (Charnes et al. 1978).

be tested. Such a test requires that a distribution be specified for the fit measure, thus allowing the reduction in its value to be assessed statistically.

A second approach to assess statistical significance is to use jackknife or bootstrap techniques to estimate a parameter's standard deviation or its distribution. Once these are known, significance tests can be conducted. To date, bootstrapping has been used with DEA to assess efficiency ratios (e.g., Simar and Wilson 2000). However, bootstrapping has not been used to assess the relevance of a particular input or attribute weight.

Attribute Weights

Reduction in Fit. Fit in LINMAP corresponds to the sum of the one-sided error terms e_{ik} over all M paired preference comparisons. However, little is known about this fit measure C^* . Fortunately, LINMAP also reports the proportion of correctly estimated paired comparisons p , which is highly correlated with C^* . This proportion is appealing because it directly measures the ability to predict the preference information actually provided by the consumer (the stated paired comparisons).

To assess the loss in fit caused by the deletion of an attribute, we test whether the proportion of correctly estimated paired comparisons for the restricted $J - 1$ attribute model (p_r) is different from that of the full J attribute model (p_f). Using the standard test for the difference of two proportions, the test statistic for the null hypothesis that $p_f = p_r$ (i.e., $w_j = 0$) is⁴

PR-test:

$$\text{reject } H_0: w_j = 0$$

$$\text{if PR} = \left| \frac{p_f - p_r}{\sqrt{((p_f + p_r)/2)(1 - (p_f + p_r)/2)2/M}} \right| \geq z_{\alpha/2}. \quad (3)$$

Computational Methods. An alternative idea is to directly estimate the standard deviations of the estimated attribute weights using a jackknife approach. With an estimate of a parameter's standard deviation,

⁴ Because the number of paired comparisons M is large even if the number of brands N is fairly small, the normal approximation is used. A minimum of 10 brands typically are used leading to a minimum of 45 paired comparisons. Note that (3) assumes p_f and p_r are uncorrelated. This is done because the covariance of p_f and p_r is unknown and not estimable unless jackknife or bootstrap techniques are used. To do so undermines a key aim of the test—ease of calculation. Furthermore, because this covariance is almost surely positive, the denominator of (3) is actually overstated. Thus, if PR performs well, a more complex test using an estimate of the covariance should have even greater power.

a simple t - or z -based statistical significance test is possible (Efron and Tibshirani 1993). For each brand i , the jackknife procedure performs a LINMAP estimation using only the data pertaining to the other $N - 1$ brands (i.e., using $M - (N - 1)$ paired comparisons). The resulting jackknife replication attribute weight estimates $\hat{w}_j(i)$ are used to estimate the standard deviations of the estimated weights \hat{w}_j . That is,

$$s_j^{JKW} = \sqrt{\frac{N-1}{N} \left[\sum_{i=1}^N \hat{w}_j(i)^2 - \frac{1}{N} \left(\sum_{i=1}^N \hat{w}_j(i) \right)^2 \right]}.$$

Hence, a test of $w_j = 0$ is

JK-test:

$$\begin{aligned} &\text{reject } H_0: w_j = 0 \\ &\text{if } JK_j = \left| \frac{\hat{w}_j - 0}{s_j^{JKW}} \right| \\ &\quad \geq z_{\alpha/2} \quad (\text{or } t_{N-j, \alpha/2} \text{ for smaller } N). \quad (4) \end{aligned}$$

A second computational approach is the bootstrap (Efron and Tibshirani 1993). Bootstrap procedures generate information concerning a parameter's standard deviation or distribution through repeated estimations using replicated samples. We "bootstrap on the data," where each replicated sample contains M paired comparisons drawn randomly with replacement from the original M pairs. These pairs then are used as input into problem (1). Each replication thus utilizes slightly perturbed data and produces a bootstrap replication of the attribute weight estimates $\hat{w}_j(b)$. The standard deviation s_j^{BT} of these $\hat{w}_j(b)$ provides the bootstrap estimate of the standard deviation for \hat{w}_j , where

$$s_j^{BT} = \sqrt{\frac{1}{B-1} \left[\sum_{b=1}^B \hat{w}_j(b)^2 - \frac{1}{B} \left(\sum_{b=1}^B \hat{w}_j(b) \right)^2 \right]}$$

and B refers to the number of bootstrap replications. Consequently, a normally distributed test for $\hat{w}_j = 0$ is

BT-test:

$$\text{reject } H_0: \hat{w}_j = 0 \quad \text{if } BT_j = \left| \frac{\hat{w}_j - 0}{s_j^{BT}} \right| \geq z_{\alpha/2}. \quad (5)$$

The BT-statistic rests on a normality assumption concerning the attribute weights that is unrealistic. Consequently, two alternative bootstrapping procedures that allow for nonnormality are often used. Both the "percentile" and the "bias corrected and accelerated" methods directly assess a parameter's distribution rather than its standard deviation and, in so doing, generate a confidence interval that is used to assess statistical significance. We refer the reader to

Efron and Tibshirani (1993) for details on these more accurate but complex and computationally demanding methods. Because simulation results for these confidence interval-based tests are only marginally better than those for the BT-test, we present simulation results only for the BT-test.

Production Coefficients

Reduction in Fit. In the efficient frontier context, a fit measure is the correlation between the actual outputs of the firms Q_i and their estimated maximum outputs \hat{Q}_i . To assess a loss in fit, we thus need to test whether Pearson's correlation coefficient for the restricted $J - 1$ input model (r_r) differs from the correlation obtained from an estimation where the full set of inputs is used (r_f). To do so, we use Fisher's $z = \tanh^{-1}(r) = \frac{1}{2} \ln[(1+r)/(1-r)]$ because its distribution approaches normality more quickly than does the distribution of r . It follows that the test statistic for a difference between r_f and r_r is the same as that for the difference between the full and restricted models' respective Fisher's z s⁵

FZ-test:

$$\begin{aligned} &\text{reject } H_0: v_j = 0 \\ &\text{if } FZ = \left| \frac{\frac{1}{2} \ln[(1+r_f)/(1-r_f)] - \frac{1}{2} \ln[(1+r_r)/(1-r_r)]}{\sqrt{2/(N-3)}} \right| \\ &\quad \geq z_{\alpha/2}. \quad (6) \end{aligned}$$

Computational Methods. A jackknife-based test akin to that used for attribute weight significance is applicable to the production coefficients. An estimate of \hat{v}_j 's standard deviation based on the jackknife replication coefficient estimates $\hat{v}_j(i)$ allows a simple test:

JK-test:

$$\begin{aligned} &\text{reject } H_0: v_j = 0 \\ &\text{if } JK_j = \left| \frac{\hat{v}_j - 0}{\sqrt{(N-1)/N \left[\sum_{i=1}^N \hat{v}_j(i)^2 - \frac{1}{N} \left(\sum_{i=1}^N \hat{v}_j(i) \right)^2 \right]}} \right| \\ &\quad \geq z_{\alpha/2}. \quad (7) \end{aligned}$$

For a bootstrap-based test, we follow previous DEA bootstrapping papers and "bootstrap the residuals" (Simar and Wilson 2000). However, we use the bootstrap to evaluate the production coefficients rather than the efficiency ratios (Horsky and Nelson

⁵ As with the PR-test, the covariance of r_f and r_r is omitted from the denominator of (6).

1996). First, the production coefficients \hat{v}_j are estimated using LP problem (2). These coefficient estimates along with the estimated error terms, e_i , are then used to generate bootstrap samples. Each bootstrap sample is generated as follows. From the estimated frontier output for each firm, an error term is subtracted. This error term $e_i(b)$ is selected randomly with replacement from the e_i .⁶ The resulting $Q_i(b) = \sum_{j=1}^J \hat{v}_j c_{ij} - e_i(b)$ are then associated with the c_{ij} via problem (2), which generates a bootstrap replication of the coefficient estimates $\hat{v}_j(b)$. As with the attribute weights, the standard deviation of the replications $\hat{v}_j(b)$ provides a bootstrap estimate of \hat{v}_j 's standard deviation. Thus, a test for $v_j=0$ is

BT-test:

$$\text{reject } H_0: v_j=0 \text{ if } BT_j = \left| \frac{\hat{v}_j - 0}{s_j^{\text{BT}}} \right| \geq z_{\alpha/2}. \quad (8)$$

Efficiency Ratio Method. For comparative purposes, we also present an efficiency ratio-based test developed by Banker (1996) that can be used to test DEA parameter significance. This test compares the sum of the efficiency ratios over the N firms for the full J input model with the sum of the efficiency ratios for the restricted $J-1$ input model. However, it requires that a distributional assumption be made concerning the efficiency ratios. This clashes with the distribution-free error framework that is inherent to DEA, and begs the question of why not use easily implemented maximum likelihood procedures if such a distributional assumption is made. For the exponential distribution assumption used in our simulation analysis, the F -test for the hypothesis that v_j equals zero is⁷

EX-test:

$$\text{reject } H_0: v_j=0 \text{ if } EX_j = \frac{\sum_{i=1}^N \theta_i^{\text{restricted}}}{\sum_{i=1}^N \theta_i^{\text{full}}} \geq F_{N, N, \alpha/2}. \quad (9)$$

Simulation Studies

Simulation studies were conducted to examine the performance of the significance tests developed above. In each simulation, utility (production output) values for a set of brands (firms) are generated from a combination of attribute weights (production coefficients) and the levels of their respective attributes (inputs) along with an error term. We then examine the ability of each test to detect significant parameters

as indeed significantly different from zero—a lack of Type I error—and detect insignificant parameters as not—the extent of Type II error.

Errors are added to the utility (production) values to incorporate phenomena such as mental errors and productivity differences. This also allows us to evaluate the robustness of each test. A test is not very useful if it does not perform well at commonly experienced error levels. However, test performance should diminish as the error increases. If not, the test is intuitively suspect. Also, as with regression t -tests, we expect test performance to increase as the number of observations relative to the number of parameter estimates (i.e., degrees of freedom) increases.

Attribute Weights

The first simulation study examines the ability of the proportion test statistic PR, the jackknife statistic JK, and the bootstrap statistic BT to detect significant and insignificant attributes in the LINMAP context. This study examines simulated data relating to a consumer's preferences toward N brands, each defined by four attributes. Attribute levels b_{ij} are generated from a uniform 0 to 7 distribution. The true attribute weights for attributes 2, 3, and 4 are set equal to 0.33, while $w_1=0$. The assumption is that the three large attribute weights are statistically significantly different from zero, while the zero weight is not. For each simulated individual, their true utility for each brand i is $\tilde{U}_i = \sum_{j=1}^4 w_j b_{ij}$. Adding an error term to each brand's true utility generates the stated utility levels, $U_i = \tilde{U}_i + \varepsilon_i$. The error terms ε_i are distributed normally with a mean of zero and a standard deviation of $(\lambda/2)\bar{\tilde{U}}$, where λ is a constant and $\bar{\tilde{U}} = (1/N)\sum_{i=1}^N \tilde{U}_i$.⁸ The brand with the highest stated utility is said to be the most preferred and so on. The M paired comparisons LINMAP uses to estimate the attribute weights \hat{w} of problem (1) are derived from this ranking.

Twenty-five simulated cases were run comprising five error levels, $\lambda=0.1, 0.2, 0.3, 0.4, 0.5$, and five brand quantities, $N=10, 15, 20, 25, 30$. The lower values for λ ($\lambda \leq 0.2$) correspond to typically reported error levels (Horsky and Rao 1984). The higher λ are used to assess the performance of the statistics at unusually large error levels.⁹ The levels for N reflect the number of brands used in multiattribute surveys. Generally, 10–20 brands are used in surveys

⁸ This puts roughly 95% of the error terms within two standard deviations ($\lambda\bar{\tilde{U}}$) of \tilde{U}_i . Uniform or half-normal distributions for ε_i generate similar simulation results.

⁹ To assess the magnitude of the error levels, we provide two statistics. The correlation between the stated and true preference rankings on average decreases almost linearly from 0.99 for $\lambda=0.1$ to 0.78 for $\lambda=0.5$, regardless of N . Similarly, the percentage of paired comparisons that are stated correctly decreases almost linearly from 0.95 to 0.80.

⁶ A slightly modified bootstrap suggested by Simar and Wilson (2000) actually is used. This process randomly perturbs the bootstrapped residuals to make the distributions of the parameter estimates more continuous. Test results using the unperturbed process described in the text differ little from those of this modified process.

⁷ A test based on the half-normal distribution performs similarly.

Table 1 LINMAP Simulation Results for True Nonzero Attribute Weights: Percentage Correctly Identified as Statistically Significant (5% Level)^a

Number of brands <i>N</i>	10					15					20					25					30									
	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5					
Statistical tests																														
PR	94	86	79	71	67	99	97	87	80	72	100	98	94	85	82	100	100	97	92	87	100	100	98	97	94	100	100	98	97	91
JK	68	49	43	36	31	97	92	82	73	64	100	100	96	91	87	100	100	98	97	94	100	100	100	100	100	100	100	100	100	100
BT ^b	95	91	87	81	78	100	100	99	98	96	100	100	100	99	99	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Regression <i>t</i> -statistic	96	88	79	64	55	100	99	95	87	74	100	100	99	92	87	100	100	100	99	91	100	100	100	100	100	100	100	100	100	99

^aPercentage of attribute weights with true value equal to 0.33 for which the statistical significance of the test statistic is found to exceed 5% (one sided), and therefore are correctly identified as significantly different from zero. This percentage equals 1—Type I error.

^b200 bootstrap resamples are used to calculate the BT-statistic.

that collect rank-order preference data on real brands. A greater number of hypothetical concepts are often used in conjoint studies. For each of the 25 cases, 200 simulated individuals were generated. Consequently, to assess how well the fit-based correlation test PR identifies significant attribute weights as significant a total of 600 one-sided tests of whether $w_j = 0, j = 2, 3, 4$ were carried out. A like number of tests were carried out using the JK- and BT-statistics. Similarly, to assess Type II error—the propensity to misidentify insignificant weights as significant—200 two-sided tests of whether $w_1 = 0$ were carried out for each statistic.

Table 1 presents the simulation results concerning how well these three tests detect significant attribute weights to be statistically significant (5% level). For comparison purposes, we also report *t*-statistic results from a regression of the brand ranks against their attribute levels. When the degrees of freedom are 16 or more ($N \geq 20$), all three tests as well as the regression *t*-statistic have good detection capabilities even at high error levels ($\lambda > 0.2$). Because the number of concepts used in conjoint studies is of this magnitude, there is no problem in studies of this nature. One must be more cautious in studies using real brands where typically $N < 20$. When $N = 15$ at reasonable error levels ($\lambda \leq 0.2$), all four tests again show a strong ability to discern significant weights. However, at $N = 10$, performance falls off for all the tests, but especially for the JK-test. This relatively poor performance at low degrees of freedom ($10 - 4 = 6$) is consistent with previous findings. Horsky and Rao (1984) find that about eight degrees of freedom are needed for LINMAP estimators to be close to the true parameters. Klahr (1969) reports that a similar number of degrees of freedom are needed to reliably identify the attribute space when performing a nonmetric multidimensional scaling analysis of similarities. Banker et al. (1984) find that the number of firms should be three times the number of inputs plus outputs to preserve the discriminating power of DEA models.¹⁰

¹⁰ The regression *t*-statistic identifies significant attributes well, but, as discussed below, has a greater tendency toward Type II error. In

Results pertaining to the tests' abilities to identify insignificant attributes as insignificant are presented in Table 2. The PR-statistic proves extremely adept at recognizing insignificant parameters as insignificant. The JK-statistic also performs well, but not to the level of PR, regardless of the error level or degrees of freedom. Alternatively, the BT-test performs very poorly. Regression *t*-statistics also perform quite poorly if $N < 20$ and never perform as well as PR.

Production Coefficients

A second simulation study examines the performance of the fit-based FZ, the jackknife-based JK, and the bootstrap-based BT-statistics. Simulated data relating to the production levels of N firms, each defined by four inputs, are derived. The levels of the inputs c_{ij} are generated from a uniform 0 to 10 distribution. The production coefficients $v_j, j = 2, 3, 4$, are set equal to one, while $v_1 = 0$. Subtracting an error term from each true maximum output $\tilde{Q}_i = \sum_{j=1}^4 v_j c_{ij}$ results in actual production levels that reflect production inefficiencies, $Q_i = \tilde{Q}_i - \varepsilon_i$. The error terms ε_i are distributed uniformly with a range of zero to $\lambda \tilde{Q}_i$, where λ is a constant and $\tilde{Q} = (1/N) \sum_{i=1}^N \tilde{Q}_i$.¹¹

Twenty-five simulated cases were run using five error levels, $\lambda = 0.1, 0.2, 0.3, 0.4, 0.5$, and five industry sizes, $N = 20, 30, 40, 50, 100$. The highest error levels analyzed exceed the inefficiencies ($\lambda < 0.4$) generally found (e.g., Charnes et al. 1988, Horsky and Nelson 1996).¹² Previous studies generally utilize 30 or more observations. We also included $N = 20$ and $N = 100$ to see the impact of degrees of freedom. For each of

addition, simulation analysis finds that LINMAP attribute weight estimates are closer to the "true" weights than are regression estimates, as measured by mean absolute deviation, mean standard error, and correlation.

¹¹ Similar results are obtained utilizing a half-normal distribution for ε_i .

¹² In our simulated data, the average "true" efficiency ratio relating the actual and true maximum production quantities, Q_i and \tilde{Q}_i , increases in a nearly linear manner from 0.06 for $\lambda = 0.1$ to 0.30 for $\lambda = 0.5$, regardless of N . Given the use of the uniform distribution, the maximum inefficiency approaches λ .

Table 2 LINMAP Simulation Results for Attribute Weights with True Weight of Zero: Percentage Incorrectly Identified as Statistically Significant (5% Level)^a

Number of brands <i>N</i>	10					15					20					25					30				
Error level λ	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
Statistical tests																									
PR	0	3	2	3	2	2	1	2	2	2	1	1	1	2	0	0	0	0	0	0	0	0	0	0	0
JK	0	1	3	2	4	1	2	4	5	5	4	3	5	7	9	1	3	6	8	12	1	3	5	9	8
BT ^b	1	10	11	13	19	7	15	18	23	24	11	19	23	34	37	8	21	32	46	49	9	19	33	46	44
Regression <i>t</i> statistic	9	15	10	12	10	7	7	10	8	9	7	10	8	11	6	8	6	7	4	11	5	7	4	7	4

^aPercentage of attribute weights with true value equal to zero for which the statistical significance of the test statistic is found to exceed 5% (two sided), and therefore are incorrectly identified as significantly different from zero. This percentage constitutes Type II error.
^b200 bootstrap resamples are used to calculate the BT-statistic.

the 25 cases, 300 hypothetical industries were generated and the LP problem (2) estimation of $\hat{Q}_i = \sum_{j=1}^J \hat{v}_j c_{ij}$ carried out. Then, FZ-, JK-, and BT-tests of the null hypothesis that $v_j=0, j=1,2,3,4$ were carried out. For comparative purposes, Banker's (1996) EX-test also was evaluated.

Simulation results are provided in Tables 3 and 4. The entries in each cell of Table 3 represent the percentage of the 900 coefficients with a true value of one for which the null hypothesis that $v_j=0$ is correctly rejected at the 5% level. The FZ- and JK-tests perform very well at correctly identifying significant coefficients as statistically significant unless both the error level is high ($\lambda \geq 0.4$) and the number of observations is low ($N \leq 30$). The BT-test performs very well at all error and brand levels and, in fact, outperforms both the FZ- and JK-tests in every case. Furthermore, in all but two cases ($\lambda \geq 0.4$ and $N=20$), the FZ-test outshines the JK-test. In addition, for all N and λ , the BT-, JK-, and FZ-tests all outperform the EX-test, and the magnitude of this superiority increases as the error λ increases.

Table 4 presents the percentage of input 1 coefficients misidentified as statistically significant. While the bootstrap BT-test does a great job of identifying significant coefficients as significant, Table 4 unfortunately shows that it also has a strong tendency to misidentify insignificant coefficients as statistically significant. On the other hand, FZ and EX, and to a

lesser degree, JK, rarely misidentify insignificant coefficients as significant irrespective of the error level or the degrees of freedom.

Summary and Recommendations

To make good marketing decisions, the firm must be able to identify the attributes that are relevant to the consumer. Similarly, to properly understand their production and cost structures, the firm must be able to assess which inputs impact production and cost levels. Unfortunately, the usefulness of LP-based procedures in addressing these issues (i.e., the estimation of multiattribute utility functions and multi-input production or cost functions) is hindered by the current lack of statistical significance tests for either model's parameters. In this paper, two types of statistical tests were forwarded and examined using simulations. One style of test assesses the difference in fit between a full J variable model and a restricted $J-1$ variable model. The other uses a jackknife or bootstrap procedure to estimate the parameter's standard deviation or distribution—thereby allowing what amounts to a t - or z -test. We find that our fit-based tests and those based on the jackknife perform quite well at identifying both significant parameters as significant and insignificant parameters as insignificant.

In the multiattribute context, three tests were forwarded. The PR-test looks at the change in the proportion of paired preference comparisons that are

Table 3 Simulation Results for True Nonzero Production Coefficients: Percentage Correctly Identified as Statistically Significant (5% Level)^a

Number of DMUs <i>N</i>	20					30					40					50					100				
Error level λ	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
Statistical tests																									
FZ	100	100	98	83	65	100	100	100	98	87	100	100	100	100	98	100	100	100	100	99	100	100	100	100	
JK	100	99	95	88	80	100	100	98	95	89	100	100	98	97	95	100	100	100	99	97	100	100	100	100	
BT ^b	100	100	100	100	99	100	100	100	100	99	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
EX	100	98	90	71	58	100	98	92	82	67	100	98	92	86	71	100	99	94	87	72	100	97	96	90	78

^aPercentage of production coefficients with true value equal to one for which the statistical significance of the test statistic is found to exceed 5% (one sided), and therefore are correctly identified as significantly different from zero. This percentage equals 1—Type I error.
^b200 bootstrap resamples are used to calculate the BT-statistic.

Table 4 Simulation Results for Production Coefficients with True Weight of Zero: Percentage Incorrectly Identified as Statistically Significant (5% Level)^a

Number of DMUs <i>N</i>	20					30					40					50					100									
	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5					
Statistical tests																														
FZ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JK	4	3	4	3	4	3	1	3	3	3	3	3	1	3	3	2	2	4	3	3	3	2	4	2	3	3	2	4	2	3
BT ^b	32	17	15	17	17	24	14	14	15	18	27	19	18	13	16	21	19	18	13	14	20	16	14	15	13	20	16	14	15	13
EX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

^aPercentage of production coefficients with true value equal to zero for which the statistical significance of the test statistic is found to exceed 5% (two sided), and therefore are incorrectly identified as significantly different from zero. This percentage constitutes Type II error.

^b200 bootstrap resamples are used to calculate the BT-statistic.

correctly estimated if an attribute is deleted from the utility function. The JK-test is a simple z-test in which the standard deviation of the parameter estimate is derived utilizing a jackknife procedure. An analogous BT-test is derived using bootstrapping. With standard quality preference data, the PR- and JK-tests do very well unless the number of brands exceeds the number of parameters by six or less. This minimum level in degrees of freedom is easily exceeded in most studies. Alternatively, the BT-test suffers from substantial Type II error. Thus, because the PR-test slightly outperforms the JK-test, is simpler to understand, and requires less computation, it has the most merit in LINMAP applications.

In the production (or cost) frontier context, three tests also were formulated. The fit-based FZ-test examines the reduction in the correlation between actual and estimated maximum outputs when all inputs are modeled and when an input is deleted. Both this FZ-test and the jackknife-based JK-test perform well at identifying significant parameters at commonly found error (inefficiency) levels. At higher error levels, these two tests require additional degrees of freedom to perform well. The number of firms required for this, however, easily falls within the typical data set size of 30 or more. Both tests outperform the efficiency ratio-based EX-test. Since the fit-based FZ-statistic is more intuitive and less computer intensive to generate than the jackknife-based JK-statistic, its use is preferred. While the bootstrap-based BT-test is outstanding at identifying significant coefficients as significant, its propensity to misidentify insignificant coefficients as significant leaves it suspect.

In summary, this paper takes a first cut at developing easy, intuitive, and accurate statistical significance tests for linear programming parameters. The availability of such tests should increase the use and acceptability of LP procedures in the management contexts for which they have been found appropriate. It may also lead to more easily accepted new applications.

References

- Banker, R. 1996. Hypothesis tests using data envelopment analysis. *J. Productivity Anal.* 7 139-159.
- Banker, R., A. Charnes, W. Cooper. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Sci.* 30 1078-1092.
- Charnes, A., W. Cooper, E. Rhodes. 1978. Measuring efficiency of decision making units. *Eur. J. Oper. Res.* 2 429-444.
- Charnes, A., W. Cooper, T. Sueyoshi. 1988. A goal programming/constrained regression review of the Bell System breakup. *Management Sci.* 34 1-26.
- Efron, B., R. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Evans, D., J. Heckman. 1988. Natural monopoly and the Bell System: Response to Charnes, Cooper and Sueyoshi. *Management Sci.* 34 27-38.
- Goldfeld, S., R. Quandt. 1972. *Nonlinear Methods in Econometrics*. North-Holland Publishing Company, Amsterdam, The Netherlands.
- Hauser, J., S. Shugan. 1980. Intensity measures of consumer preference. *Oper. Res.* 28 278-320.
- Horsky, D., P. Nelson. 1996. Evaluation of salesforce size and productivity through efficient frontier benchmarking. *Marketing Sci.* 15 301-320.
- Horsky, D., M. Rao. 1984. Estimation of attribute weights from preference comparisons. *Management Sci.* 30 801-822.
- Horsky, D., P. Nelson, S. Posavac. 2004. Stating preference for the Ethereal but choosing the concrete: How the tangibility of attributes affects attribute weighting in value elicitation and choice. *J. Consumer Psych.* 14 132-140.
- Jain, A., F. Acito, N. Malhotra, V. Mahajan. 1979. A comparison of the internal validity of alternative parameter estimation methods in decompositional multiattribute preference models. *J. Marketing Res.* 16 313-322.
- Kamakura, W., R. Srivastava. 1986. An ideal-point probabilistic choice model for heterogeneous preferences. *Marketing Sci.* 5 199-219.
- Klahr, D. 1969. A Monte-Carlo investigation of the statistical significance of Kruskal's nonmetric scaling procedure. *Psychometrika* 34 319-330.
- Seiford, L., R. Thrall. 1990. Recent developments in DEA: The mathematical programming approach to frontier analysis. *J. Econometrics* 46 7-38.
- Simar, L., P. Wilson. 2000. Statistical inference in nonparametric frontier models: The state of the art. *J. Productivity Anal.* 13 49-78.
- Srinivasan, V., A. Shocker. 1973. Linear programming techniques for multidimensional analysis of preferences. *Psychometrika* 38 337-369.