TRANSPORTATION RESEARCH RECORD 874

Advances in Trip Generation and Quantitative Methods

TRANSPORTATION RESEARCH BOARD

NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY OF SCIENCES
WASHINGTON, D.C. 1982

Transportation Research Record 874
Price $6.00
Edited for TRB by Brenda J. Vumbaco

mode
1 highway transportation

subject areas
12 planning
13 forecasting

Library of Congress Cataloging in Publication Data
National Research Council. Transportation Research Board.
Advances in trip generation and quantitative methods.

(Transportation research record; 874)
1. Origin and destination traffic surveys–Congresses. 2. Trip generation–Congresses. 3. Choice of transportation–Congresses. I. National Research Council (U.S.). Transportation Research Board. II. Series.
TE7.H5 no. 874 [HE370.075] 380.5s 83-2447
ISBN 0-309-03405-1 [388.3'14'072] ISSN 0361-1981

Sponsorship of the Papers in This Transportation Research Record

GROUP 1–TRANSPORTATION SYSTEMS PLANNING AND ADMINISTRATION
Kenneth W. Heathington, University of Tennessee, chairman

Transportation Forecasting Section
George V. Wickstrom, Metropolitan Washington Council of Governments, chairman

Committee on Passenger Travel Demand Forecasting
David S. Gendell, Federal Highway Administration, chairman
Moshe E. Ben-Akiva, Daniel Brand, David J. Dunlap, Robert T. Dunphy, Raymond H. Ellis, Robert E. Gall, Thomas F. Golob, Walter G. Hansen, David T. Hartgen, Thomas J. Hillegass, Joel L. Horowitz, Stephen M. Howe, Lidia P. Kostyniuk, T. Keith Lawton, Steven Richard Lerman, Eugene J. Lessieu, Frederick A. Reid, Martin G. Richards, Gordon W. Schultz, Gordon A. Shunk, Bruce D. Spear, Anthony R. Tomazinis, Edward Weiner, Yacov Zahavi

Committee on Traveler Behavior and Values
David T. Hartgen, New York State Department of Transportation, chairman
Peter M. Allaman, Julian M. Benjamin, Katherine Patricia Burnett, Melvyn Cheslow, David Damm, Michael Florian, Thomas F. Golob, David A. Hensher, Joel L. Horowitz, Frank S. Koppelman, Steven Richard Lerman, Jordan J. Louviere, Arnim H. Meyburg, Richard M. Michaels, Alfred J. Neveu, Robert E. Paaswell, Earl R. Ruiter, Yosef Sheffi, Peter R. Stopher, Antti Talvitie, Mary Lynn Tischer

James A. Scott, Transportation Research Board staff

Sponsorship is indicated by a footnote at the end of each report. The organizational units, officers, and members are as of December 31, 1981.

Contents

Authors of the Papers in This Record

Daganzo, Carlos F., Institute of Transportation Studies, University of California, Berkeley, CA 94720
Horowitz, Joel L., Department of Geography, University of Iowa, Iowa City, IA 52242; formerly with the U.S. Environmental Protection Agency
Ou, Fong-Lieh, Gifford Pinchot National Forest, U.S. Department of Agriculture, Vancouver, WA 98662
Reid, Fred A., Consultant, 2439 Russell Street, Berkeley, CA 94705
Yu, Jason C., Department of Civil Engineering, University of Utah, Salt Lake City, UT 84112

Critique of ITE Trip Generation Rates and An Alternative Basis for Estimating New Area Traffic

FRED A. REID

The commonly used household vehicle trip rates from the Institute of Transportation Engineers' Informational Report, Trip Generation, are nearly twice as large as equivalent data from survey-based sources such as the 1977 National Personal Transportation Study and the various regional planning agencies. A number of potential reasons for the difference, including sample demographics, underreporting, and non-home-based trips, estimating the effect of each. Major factors are the atypically large incomes and family sizes of areas from which the ITE data came. The quantifiable factors explain two-thirds of the difference. Other reasons, such as the ITE rates being measured in new areas without maturity of land uses or resident's habits, are speculated as the cause of the remaining difference. Even the quantifiable reasons for differences in characteristics between a new area and the one from which source data come are often ignored in traffic estimates. Considering these factors ITE data are claimed to lead to excessive local road capacity decisions. Survey-based rates are more consistent with other national accounts and observations of travel. An alternative method for traffic estimation for new residential developments, based on available survey sources, is presented and recommended.

Traffic load estimates and road requirements for new residential developments are often based on per-household trip rates from the Institute of Transportation Engineers Informational Report, Trip Generation (1). ITE rates for the typical number of vehicle trips per day produced by single and/or corresponding multifamily housing types are multiplied by the mixes of housing planned to give the traffic expected. Other steps or refinements may adjust the basic rates for the demographics of the area, determine peak-hour trips, subdivide the estimates by zones and a road network, add non-residential and external trips, and/or apply traffic level of service standards to determine the number and sizes of roads required for the development.

Typical of the tripmaking rates in the ITE manual is the claim that the average single-family housing unit generates 10 vehicle trips per day. Generation rates from 3.7 to 6.1 are given for different densities of multifamily housing. These rates are considerably higher than reported for population averages by most national and regional survey sources.

The National Personal Transportation Study (NPTS) reports that the average U.S. household made 4.0 vehicle trips/day in 1977 (2). Characteristics of Urban Travel Demand, a compilation of travel survey data from about 40 metropolitan areas in the United States in the 1960s, averages to 5.3 home-based person trips per day (equivalent to about 4.4 vehicle trips/day between all locations) (3). The 1976 Urban and Rural Travel Survey of 7600 households in the six-county Southern California Association of Governments (SCAG) region reported an average vehicle-trip rate of 5.7/household/day (4).

Even if the ITE figures are weighted by the corresponding mixes of the housing types in the above populations, they still imply twice the national rate and 50 percent higher than reported in southern California--7.8 and 8.5 vehicle-trips/day, respectively.

There are straightforward reasons for some of these differences. Data-collection methods and demographics are different for the sources. However, as the next section of this paper shows, the ITE generation rates are still 34 percent high after adjustments are made for identifiable factors. This suggests that estimates of road requirements for new residential developments based on the ITE rates may lead to construction of excessive capacity and unnecessary expenditures. At a time when automobile travel may be leveling off or even decreasing and resources are scarce, this should be remedied.

The paper goes on to apply an alternate method of estimating traffic to an actual new community based on travel survey data. The result is compared with an independent estimate based on ITE-type data, with the latter again appearing excessively large. The conclusion speculates on the persisting part of the differences. Because of the biases implied in the ITE sampling methods, travel survey data are recommended as the basis for sizing new roads.

ADJUSTMENTS FOR DIFFERENCES BETWEEN ITE AND NPTS METHODS

Before detailing the quantifiable differences between the two sources, issues that do not contribute to their differences must be cleared away. All the data sources make a distinction between person and vehicle trips and are stated for the latter. Vehicles in the NPTS survey include, in addition to automobiles, vans, light trucks, recreational vehicles (RVs), motorcycles, and mopeds. Though the other sources do not define their inclusion of vehicles, they are unlikely to be more comprehensive. So the NPTS rates apparently include at least the vehicles in the cordon measurements. All sources define a trip as each one-way leg, so a round trip counts as two vehicle trips. Reports on trends in tripmaking rates between the early 1970s and 1977, the interval between most of the ITE and NPTS observations, do not show significant changes (2,5,6). There was a 6 percent decline on a per household basis due to greater household growth than for travel or the population. This effect is not included here, but it may have a bigger effect on future projections and is discussed later.

The initial set of adjustments below is to reconcile the obvious differences in data collection between cordon line and survey methods. Following this, adjustments are also made for the apparent atypical demographics of the ITE sample. The reconciliation procedure is to start from the average trip rate of one of the sources (NPTS) and adjust it for each quantifiable difference in data-collection method or demographics to try and represent the basis of the other source (the ITE rates averaged at the U.S. housing mix).

First, the NPTS trip rates are adjusted to the same (cordon line) basis as the ITE sources by subtracting non-home-based trips and then adding back proxy data for the trips coming to houses by nonresidents (visitors, service vehicles, etc.). National data typically report about 80 percent of trips to be home-based. In the SCAG area 70 percent are home-based. In the San Francisco region, 21 percent were non-home-based in the 1960s. The national figure of 20 percent is subtracted from the NPTS rate in the second row of Table 1 for this factor.

The SCAG data and data from the Washington, D.C., region indicate that non-resident trips to homes amount to 16-19 percent of total area trips (+21 to +23 percent of home-based trips) (7). The third row

of Table 1 adds 21 percent to the NPTS rate for this effect.

Another probable reason for differences is the underreporting of trips by interviewees. Two indications of the size of this effect are available. Correction factors are typically applied to survey-based models of regional tripmaking to gain agreement with ground counts. The models of the California Department of Transportation (Caltrans) LARTS Division, Los Angeles, indicate an adjustment factor of 19.6 percent between SCAG area reported trip rates and (LARTS-DTIM) production rates per household. Adjustments in the Sacramento region run from 0 for work trips to 20 percent for other home-based trips. Checks of reported miles driven against driver trips multiplied by trip distances within the NPTS data suggest about 28 percent underreporting. Omission of trips probably accounts for the bulk of this error. About 20 percent is taken to represent this underreporting, the third adjustment factor shown in Table 1. This brings the survey rates to equivalence with the cordon line basis of the ITE data.

The ITE descriptions of the average demographics of their single-family category suggest other important reasons for the trip rate difference. Homes in this sample have an average of 3.7 persons, 1.6 vehicles; are built at a density of 3.5 units/acre; are newer than the average; and, following from all these characteristics, have higher incomes. Each of these factors contributes to the ITE rates being higher than surveys of typical tripmaking.

Table 2 shows additional adjustments to the survey-based trip rate to try and make it represent a residential area with the average demographic and economic characteristics of the ITE sources. The rates are adjusted by using information on the sensitivity of tripmaking to the relative demographic characteristics of the ITE and the national samples [derived from the SCAG study (4); national demographics are from the 1976 housing census].

The first identifiable demographic difference is a NPTS sample family size of 2.92 persons/household versus 3.17 for ITE (when weighted by the U.S. housing mix). The NPTS average household incomes were $12 460/year in 1976. No direct report of incomes was given with the ITE data. However, its other demographic, automobile ownership, and sample area characteristics were very similar to Orange County, California, in 1976. The $15 400 average household income for that county was thus inferred to represent the ITE sample. Individually, these demographic differences imply a 6.4 percent and a 21 percent higher vehicle trip rate for the ITE sample for family size and incomes, respectively. Assuming a correlation factor of 0.33 between family size and income, the joint effect of these factors is a 23 percent higher trip rate. This is shown as the first adjustment in Table 2.

The ITE data were taken in new and low-density areas where transit shares of trips are insignificant. The nationwide transit share of all trips was 2.8 percent in 1977 according to NPTS. This last adjustment in Table 2 brings the survey-based trip rate to equivalence with measurements taken in areas without transit.

Table 2 still concludes with a 34 percent unexplained difference between the adjusted NPTS vehicle trip rate and that from ITE. A similar set of adjustments on the southern California trip surveys leave 21 percent of the difference between it and the ITE data unexplained. Thus, even if the ITE trip rates were applied to areas in the nation with the same average demographics as the ITE sources, traffic estimates would typically be 34 percent too high. As stated originally there would be discrepancies of nearly 100 percent if ITE rates were applied directly to areas with average national demographics.

Apparently there are significant reasons beyond the adjustments in Tables 1 and 2 for the ITE trip rates to differ from recent surveys. No adjustments were made for land use densities above; weighting of the average ITE rate by the housing type mix may not correct for density differences. No account was made for possible differences due to the maturity of the areas or the ages of their populations. Further reasons are speculated for this difference in the conclusions. Of course, all of the adjustment factors applied in the table for underreporting and other effects are subject to error.

The difference in the trip rates is large compared with likely errors in the adjustment factors. It is larger than any of the individual adjustments. Either underreporting or income adjustments would have to be more than 50 percent to account for it.

The NPTS rate is also much more consistent with national accounts of vehicle miles traveled and annual energy consumed by personal vehicles. If the 7.8 trips/day U.S. average implied by the ITE rates are multiplied by the average trip lengths, number of households, and fuel efficiency of vehicles for 1972, the annual energy consumption implied is about 17 quadrillion Btu's, also about twice the published reports of U.S. energy consumption by personal passenger cars (8).

Because the difference in the ITE data remains excessive under examination, an alternative method for estimating the traffic for a new residential development based on travel survey data is presented

Table 1. Reconciliation of NPTS to ITE trip generation rates for data-collection methods.

Source or Adjustment Factor	Adjustment (%)	Resulting Vehicle Trips per Avg Household per Day
NPTS	Baseline	4.0
Remove non-home-based trips	-20	3.2
Add visitor and service vehicle trips	+21	3.9
Add for underreporting	+20	4.6
Cumulative adjustments for cordon equivalence	+16	4.6

Table 2. Reconciliation of NPTS to ITE trip generation rates for demographic differences.

Source or Adjustment Factor	ITE Sample Parameter	NPTS Sample Parameter	Adjustment[a] (%)	Resulting Vehicle Trips per Avg Household per Day
NPTS rate adjusted to cordon basis (from Table 1)			+16	4.6
Family size	3.17/household	2.92/household	+6.4 } [b]	5.7
Income per household			+21 }	
Transit trips	2.8 percent share	Negligible	+2.8	5.8
Cumulative cordon and demographic adjustments			+45	5.8
Avg ITE vehicle trips per household at U.S. housing mix				7.8
Remaining difference ITE versus NPTS (adjusted)			34	

[a]Trip rate sensitivities to changes from Caltrans (4). [b]Correlation of 0.33 assumed between income and family size.

next. As well as being based on more realistic data for trip rates, the new method accounts for the expected demographic and economic characteristics of the area's residential market and the effects of recent fuel prices. This will be compared with a traffic estimate by a consultant for the same area plan based on ITE-type data.

TRAFFIC ESTIMATION BASED ON HOUSEHOLD TRAVEL SURVEYS

The revised estimate is for the total residential vehicle traffic generated by the 43 000 dwelling units planned for the 18 000-acre Chino Hills area of San Bernardino County, California. Only average household vehicle trip rates and the total traffic produced by the land uses in the plan will be estimated. The consultant's estimate distributed traffic over the transportation network of the area. Though a network analysis is beyond the scope of this study, a useful comparison can be made of the total trips produced by both estimates and inferences made for their proportional effect on the road network. The estimate here goes beyond the prior one in source data and demographic and economic adjustments.

Table 3. Vehicle driver trips per household by income level and housing unit type.

Household Income ($)	SH	MH	Overall
Los Angeles County			
Under 6000	3.6	1.8	2.5
6000-9999	4.6	3.4	3.9
10 000-14 999	5.9	4.8	5.4
15 000-24 999	8.4	6.6	7.9
25 000-39 999	9.7	5.6	8.8
40 000 and over	9.6	6.9	9.1
Overall	6.8	3.8	5.5
Orange County			
Under 6000	4.3	2.4	2.9
6000-9999	4.9	4.8	4.8
10 000-14 999	7.2	5.2	6.2
15 000-24 999	7.8	6.6	7.5
25 000-39 999	10.2	8.0	9.9
40 000 and over	9.2	-[a]	9.2
Overall	7.6	4.8	6.5
Riverside and San Bernardino Counties			
Under 6000	3.3	1.8	2.8
6000-9999	4.9	4.9	4.9
10 000-14 999	6.4	6.4	6.4
15 000-24 999	9.1	5.9	8.7
25 000-39 999	10.7	6.5	10.5
40 000 and over	9.3	3.0	8.8
Overall	6.2	4.1	5.8

[a] No usable data.

The SCAG/Caltrans 1976 Travel Survey discussed above is the source of the trip rates for this estimate. Table 3 from this survey shows the vehicle trips made per household by various income and housing-type groups for three counties in the Chino Hills region. The vehicle rates are for all trips in the counties (including commercial trips except heavy trucks), even though they have been expressed per household (person trips/household were typically 1.4 times higher). Rates under the headings SH are those for single-family homes, those under MH are the trips per day for multifamily housing units. The overall rates are for the 1976 mix of housing types in each county.

The steps in estimating the residential traffic rates for the Chino Hills plan are

1. Start with the rates in nearby developed areas,
2. Correct for underreporting of trips by households,
3. Adjust for expected household incomes in area,
4. Subtract non-home-based trips,
5. Add incoming visitor and service-commercial trips,
6. Adjust for expected mix of housing types, and
7. Correct for recent declines in tripmaking.

All of these adjustments are summarized in Table 4 and detailed below. The joint effect of these adjustments is taken as the product of the individual percent adjustments. Correlations between all the factors are assumed zero except income and housing type; their correlation is accounted for with an assumed coefficient of -0.5.

The baseline for the trip rate adjustments is the average traffic produced by households in the county where Chino Hills will be developed. This is seen from Table 3 to have been 5.8 vehicle trips/day in 1976. Since the development is on the boundary of three counties and similar to those in the others, the unweighted average of the daily rates in San Bernardino, Orange, and Riverside (SB/ORA/RIV) Counties is computed from the table as a more representative baseline. This is 6.15 vehicle trips/day, the first entry in Table 4. The average demographics for these counties were 2.96 persons/household, 71 percent multifamily housing, and $13 000 household income in 1976.

The first adjustment is for the underreporting characteristic of household travel surveys. As stated above the correction factor for this in SCAG area studies is 20 percent. The effect of this adjustment is an average household rate of 7.4 vehicle trips/day, as shown in the second row of the table.

The 1979 average household income for the planning area was $22 400 annually. By using the average of the tripmaking rates of households in this income category from the table for SB/ORA/RIV Coun-

Table 4. Traffic estimate for Chino Hills residential development based on Caltrans/SCAG survey data.

Source or Adjustment Factor	Study Area Parameter	Adjustment (%)	Avg Vehicle Trips per Household per Day
1976 SCAG/LARTS survey	Baseline		6.15
Addition for underreporting	All home-based	+20	7.4
Factor for			
High income per household	$22 400	+31.7	9.7
High multifamily	33 percent multifamily	-2.0	
Remove non-home-based trips	0 percent[a]	-30.7	8.3
Add visitors and service trips	100 percent	+16.4	
Factor for fuel prices and economy	1980 conditions	-4.5[b]	7.9
Avg vehicle trip rate for all households			7.9
Total vehicle trips per day (43 391 housing units)			343 200

[a] Added separately later in traffic forecast. [b] From Travel and Related Factors in California (6).

ties indicates that they produce 32 percent more trips than the average. The income adjustment is taken here to include the effect of automobile ownership, which also influences tripmaking but is highly correlated to income.

In 1976 multifamily housing made up 28.6 percent of the stock in SB/ORA/RIV Counties. Assuming that the proportion of Chino Hills planned for 12 or more housing units per acre will all be multifamily, the area will have 33 percent multifamily units at full buildout. Caltrans survey tabulations on the dependence of traffic generation on housing type indicate that such a housing mix produces 2 percent fewer vehicle trips than the existing mix. The net effect of this and the correlated income factor is a 31 percent increase in the trips, or 9.7 vehicle trips per household per day. This effect is shown in the fourth entry of Table 4.

No adjustment is made for the densities planned for Chino Hills. At the gross area scale they are approximately equal to the existing developed areas from which the trip rates came. This contrasts with developments characterized by ITE that may be at half the prevailing densities. Sensitivities of vehicle trip rates to densities imply the ITE sample may give 15 percent more trips.

The Caltrans survey shows that non-home-based travel--for example, between work and/or personal business destinations--in San Bernardino County is 30.7 percent of the above figures. Thus residential-based traffic is that much lower. However, the traffic to residences by visitors, service vehicles, etc., must be added. Caltrans models for trip attraction show this to be 16.4 percent of all trips (from residential weight of "other-to-home" attractions, according to Gerald Bare, Caltrans District 7, LARTS Division, Los Angeles). The net effect of these two factors is a reduction of 14.3 percent from regional trips. This is shown as the fifth and sixth entries in Table 4.

Studies have shown that trip rates have been falling off more persistently in the late 1970s (5,6,9). The recent Caltrans report shows that vehicle miles traveled per household decreased 3.6 percent between 1978 and 1980, while vehicle trips per household went down 12.6 percent. By using the Caltrans rates of change, adjusted by the NPTS ratios of miles to trips, a decrease of 4.5 percent in vehicle trips per household is implied for California between 1976 and 1980. This adjustment is shown in the seventh entry of Table 4. Since fuel prices are likely to increase disproportionately with the rest of the economy in the future, this may only be the beginning of a continuing trend reducing travel and therefore requirements for road capacity.

The net effect of these corrections and adjustments indicates that the average vehicle trip rate from all housing in Chino Hills will be 7.9 vehicle trips/day. At this rate the 43 000 residences in the Chino Hills plan would generate 343 000 vehicle trips/day. A corresponding estimate made by the traffic consultant for the Chino Hills Plan was 447 000 vehicle trips/day to and from residences, or an overall average of 10.3 per household (10). This is 30 percent above the estimate based on the adjusted Caltrans survey data.

CONCLUSIONS

There is a significant difference between the new estimate and that based on ITE-type data. If the demographic-economic characteristics of the case study area were closer to the southern California average, the difference could have been 50 percent. The reasons for the difference appear to be that ITE household trip rates are from samples with the following special residential characteristics: (a) unusually high family sizes and incomes, (b) unusually low development densities, (c) pre-fuel-crisis observations, (d) little land use mix or balance of services, and (e) lack of maturity of community activities, their populations, and trip patterns. The first two simply represent the kind of identifiable differences in samples treated in Tables 1 and 2. The others are speculated reasons for the remaining 21 percent difference for the equivalent southern California comparison.

Area survey data cover land uses that have a development maturity more appropriate for long-range planning. The ITE data come from limited, exclusive use new areas in the late 1960s and early 1970s. They had not reached the maturity of supporting services, population age distribution, and activity habits of the residents that occur with the long-term fill out of the area--the maximum use condition for which traffic and roads need to be sized. New areas may not be established as a place to live and the new residents may have greater linkages to prior residences or the region as a whole. Some 56 percent of trips from residences are outside of local communities (more than 5 miles) and are influenced mainly by the regional context of travel. New residential developments have in the past been in regions with sparse additional uses because of land costs. They rarely had any significant travel mode alternatives for reaching desired activities. They avoided a mix of nonresidential land uses.

ITE sources usually excluded mixed land use areas because measurements on them would not be pure data for one use. But this fails to reflect the traffic reduction due to walking and short trips. Pure use data will overestimate traffic for larger developments. Current land developments are appropriately not the exclusive land-intensive residential suburbs of the 1960s. They provide a balance of residential and supporting land uses, significant transportation alternatives, and an attractive residential, shopping, and recreational environment to capture many trips. Pure use measurements may be an indication of the traffic on the most local streets, or for very small developments. Nationally, 10 percent of vehicle trips are under 0.5 mile.

Some of the reasons for the lower traffic estimate here may actually further reduce travel in the future. These are the recent declines in vehicle use due to fuel and other automobile prices, the trend toward greater housing density, vehicle leasing or renting, the growth of regional transit, the generally increasing age distribution of the population, and trends in decentralized work places and telecommuting. Thus the overestimates from ITE-type data suggested here for 1980 may become even larger in the future.

It might be argued that the trip rates used for new traffic estimates should be more liberal to account for future growth of household tripmaking or increases in housing densities. This may be a consideration, but if so it should be more explicit. With the present flattening or even decreasing of energy supplies and personal incomes, the saturation and falling off of travel rates in the 1970s may be the future pattern. Densification is more plausible for the future. However, with the price of housing and scarcity of capital, it may not be appropriate to burden new developments with such long-range speculation.

Travel survey data for an area similar to the one to be estimated are the basis of the recommended estimate above. The results here suggest that this would generally be a better basis for new area traffic forecasts than the ITE manual. Most metropolitan areas have travel survey data, by county or

smaller areas, updated for current years. The transportation demand compilations of the Urban Mass Transportation Administration mentioned earlier are also a better basis for estimation. Even where these are lacking, choosing comparable survey data and adjusting them by the methods presented here would be preferable to accepting the probable inflation of the ITE data. (Use of the ITE data also assumes a transferability of its sources to areas for which they may not have been collected.)

In times of major fuel price increases and shortages, even people in exclusively suburban settings may not continue to greatly exceed average regional travel. As shown above, people's trip rates have been coming down since the ITE data were compiled. If we do not want excessive expenditures for transportation or unnecessary encouragement of travel by overcapacity, we should not be sizing our roads from traffic generation rates taken from pre-fuel-crisis suburbs.

The lower traffic estimate here suggests that road building for a new area based on the more traditional data would be an excessive use of resources--capital, energy, and land. It would produce a road system compatible with a much more intensive land use than intended by the plan. If uses were held at the intended plan maximums, the excess road capacity would give an over-automobile-oriented character to the area that encourages excessive automobile use and defeats other aspects of the plan to encourage transportation alternatives and stop the propagation of the old syndrome of more roads, more travel, more congestion, and more energy consumption.

REFERENCES

1. Trip Generation, An Institute of Transportation Engineers Information Report. 2nd ed. ITE, Tech. Council, Committee 6A-6, Washington, DC, 1979.
2. Purposes of Vehicle Trips and Travel, National Personal Transportation Study, Report 3. FHWA, 1980.
3. Characteristics of Urban Transportation Demand: Handbook. UMTA, Rept. UMTA-IT-06-0049-78-1, 1979.
4. 1976 Urban and Rural Travel Survey: Volume 4, Summary of Findings--Travel Data. California Department of Transportation; Southern California Association of Governments, Los Angeles, 1979.
5. W.C. Lee. The Demand for Travel and the Gasoline Crisis. Motor Vehicle Manufacturers' Association, Detroit, MI, 1980.
6. Travel and Related Factors in California. California Department of Transportation, Sacramento, 1981.
7. Computer Programs for Urban Transportaton Planning, PLANPAC/BACKPAC General Information. FHWA, 1977.
8. National Transportation Statistics, Annual Report. Transportation Systems Center, U.S. Department of Transportation, Cambridge, MA, 1980.
9. Highway Travel Trends in the 1970s. FHWA, 1980.
10. W. Kunzman. Chino Hills Traffic Study of Selected Concept Plan. Kunzman and Associates, Irvine, CA, draft report, 1980.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Effect of Urban Character on Transferability of Travel Demand Models

FONG-LIEH OU AND JASON C. YU

The effect of urban character on travel demand model transferability is investigated. Urban character was described by urban area size, type of activity concentration, and geographic characteristics, while travel demand measures consisted of trip frequency, trip length, and mode choice and were grouped by work and nonwork trip purposes and by metropolitan and nonmetropolitan areas. A generalized regression dummy variable approach was used for the analysis. The study results show that urban character bears significant influence on travel demand, and the model transferability varies with demand measures and model specification. The specific findings include (a) the type of activity concentration has a significant impact on trip frequency for nonmetropolitan areas and trip length for both metropolitan and nonmetropolitan areas; (b) the influence of urban area size on mode choice and trip length is significant for metropolitan areas; (c) the impact of geographic characteristics on travel demand can be ranked in order of trip length, trip frequency, and mode choice; (d) metropolitan trip frequency models are more transferable than their nonmetropolitan area counterparts, while the transferability of trip length models of both metropolitan and nonmetropolitan areas is very low; and (e) in nonmetropolitan areas, nonwork trip frequency models are more transferable than work trip models.

Experience in modeling indicates that a powerful organizing paradigm seems to generate its own problems. This is particularly true when dealing with such a complex reality as travel behavior in which influential elements observed may be merely partial and distorted. Some of the perceived elements are rigorously analyzed while others may have been completely overlooked. For instance, in the conventional method of trip generation analysis, land use variables have been used to determine trip production and attraction. However, these variables should not be considered as all-inclusive parameters that affect travel behavior. To illustrate, if the values of significant land use variables were identical for different urban areas, habitual travel behavior would indicate that each urban area would always remain different from that of every other. This implies that some other uncertainties must exist that also influence the desire for travel.

Each urban area has its own character, which may be typified by urban area size, activity concentration, geographic characteristics, etc. The urban character can be considered as the base conditions of a given urban area that dictate the activity sys-

tem, the transportation system, and the traffic flow pattern.

If a demand model is developed for an area and is intended to be used for predicting travel demand in that area only, the effects of urban character may be neglected from analysis without jeopardizing the predictive accuracy. However, if a transferable model is desired for application of travel prediction in different areas, the distinction of urban character should become a significant factor in influencing travel behavior. The reason is that the observed travel demand patterns inherit the nature of urban character that is obviously different from one area to another.

The primary objective of this study was to identify urban character in terms of urban size, activity concentration, and geographic characteristics that affect the transferability of work and nonwork travel demand models. The significance of urban character was determined by the first set of data and then validated by a second set of data. In addition, the influence of model specification on transferability was also investigated. Travel demand measures include trip frequency, trip length, and mode choice. The main reason for choosing these measures is that composite demand measures, such as person miles of travel (PMT), person hours of travel (PHT), vehicle miles of travel (VMT), and vehicle hours of travel (VHT) can be derived. These composite demand measures have been applied as indicators of areawide transportation system performance and resource allocation.

Various approaches have been taken to identify urban character related to the transferability of travel demand models. The first approach is to classify urban areas based on the dominant economic activity (1-3). The second approach groups cities based on population and automobile availability (4). The third approach categorizes cities in accordance with their types of activity concentration (5). The fourth approach classifies cities by using factor and cluster analysis techniques (6,7). Although no study has yet been conducted for the use of all these notions in one context so that the joint effect of urban character can be investigated, this study was intended to do just that. In addition, there are obvious weaknesses in past studies to relate urban character to various types of models for transferability purposes. For instance, use of the relationship of automobile availability and trip frequency per person to urban areas may lead to the incorrect expectation that a greater automobile availability would yield a greater number of trips. As a matter of fact, cities that have higher automobile ownership rates often generated lower trip frequencies (8). Another example is that the classification based on the type of activity concentration may be more applicable to trip distribution than trip generation. The type of activity concentration has a better relationship with gravity-model friction factors than trip frequencies (9). Without the support of statistical evidence, however, these suggestions are not convincing. This study aimed to examine the applicability of the aforementioned urban classifications by using statistical analysis.

INITIAL SELECTION OF VARIABLES

A generalized regression dummy variable approach was used as the basic model form in this study. The proposed travel demand model consists of dependent and independent variables discussed in the following sections.

Dependent Variables

In this study, measures including travel demand necessary to estimate total person trips, PMT, and PHT were selected as the demand variables. They are trip generation, trip length, and mode choice in terms of trip frequency, trip duration, travel distance, percentage of travel by automobile, transit, and walk. In addition, automobile occupancy was considered because it would allow the estimation of modal-split between the automobile driver and the automobile passenger based on the percentage of travel by automobile.

In most urban transportation studies, travel demand was differentiated by work and nonwork trips. Thus person trips and trip length were categorized into work and nonwork trips. Because trip frequency is a common measure for both household models and zonal models, it was used as a measure of trip production. Mode choice models cover modes of automobile driver, automobile passenger, transit, walk, and other modes of transportation considered in this study.

Independent Variables

Conventional travel demand analysis assumes that the number of trips within an area depends on the land use of the area. Because the area's population, socioeconomic conditions, land use development, and employment reflect the land use, these activity system variables are usually used as the parameters for travel demand. In this study, relationships among land use indicators, urban character, and tripmaking activity were estimated so that the impact of changes in socioeconomic and physical environments on changes in travel demand could be statistically analyzed.

Urban Character Variables

Considerable theoretical analyses and empirical research efforts have been made to determine the factors that most influence travel behavior at the individual, zonal, and areawide levels (10-12). This study used the results of these efforts as the basic information for the selection of urban character variables.

The three independent variables considered to reflect the urban character are urban size class, activity concentration, and geographic cluster (i.e., groups of cities that share similar geographic characteristics). The urban size class contains five urban groups with a population of less than 50 000; 50 000 to 99 999; 100 000 to 249 999; 250 000 to 750 000; and more than 750 000. This classification scheme is consistent with the overall transportation policy framework of the nation. Section 134 of the Federal-Aid Highway Act of 1962, for example, specifies that urban areas with a population of 50 000 or more have comprehensive, cooperative, and continuing (3C) planning programs under way. In many local planning contexts, this legislative action draws a sharp distinction between cities of different sizes. Thus cities with population of 50 000 or more form a logical group because of their common planning guidelines, while those below 50 000 constitute a second grouping.

Next, in many demographic and economic studies, the population size of 100 000 has often been regarded as a benchmark for distinguishing urban areas with respect to social structure and economic performance. Because both factors have profound impact on travel demand, this study proposed that urban areas with a population of less than 100 000 and urban areas with a population of 100 000 or more should be divided into two different groups. Urban areas within each group share common economic characteristics.

The third distinction was made for urbanized areas with a population of 250 000 or more, most of which have performed multimodal planning programs including long-range transit planning. Thus, it would appear that urban areas with a population of 250 000 or more constitute one grouping, while those less than 250 000 are referred to as another.

Finally, a recent study indicated that a population of 750 000 is the minimum size for a city to provide certain types of transit service such as light rail (13). Urban areas falling into this category could generate enough activity to effectively use and financially support light rail facilities and service. Other studies also showed that travel behavior in large cities is significantly different from that in smaller urban areas (14). Large cities, for example, report a greater travel demand in terms of VMT on a per capita basis (15). In accordance with these research results, urban areas can be categorized into two groups--cities with population more than 750 000 and cities with 750 000 people or less.

The urban activity concentration includes the core-concentrated and the multinucleated (5,16). The reason for this distinction is that the distribution patterns of population and employment have significant bearing on travel demand. Obviously, the greater dispersion of population and employment, the higher the travel demand in terms of VMT. On the other hand, the type of activity concentration can be expressed by density. A high urban density tends to favor public transportation and depresses automobile use due to the limitation of parking space and higher operating costs.

The geographic cluster consists of nine different urban groups classified by Golob and others (6). The area characteristics used for the classification of 80 metropolitan areas included 53 variables related to arterial transportation requirements such as population, demographics, and socioeconomic measures, land use and economic activity measures, geographic factors, and travel mobility and accessibility measures. A principal components factor analysis was performed on the correlation matrix of the 53 variables (17). Based on eigenvalues, 15 orthogonal latent factors were derived. Because the 15 latent factors include most aspects of urban environment and particularly reflect geographic influence, the urban classification based on these factors was selected as representative of the geographic variables in this study.

The city grouping shows that the first two of nine homogeneous groups reflect the uniqueness and dominance in the urban hierarchy of first, New York, and second, Chicago and Los Angeles. The third group consists of large northeastern cities characterized by high residential density and transit orientation, such as Baltimore, Boston, Pittsburgh, etc. The fourth group consists of southern cities including Atlanta, Charlotte, Memphis, etc. The cities in this group have high residential density and low income. The fifth group contains midwestern cities such as Denver, Indianapolis, Oklahoma City, etc. The cities in this group are characterized by average industrialization and personal income, and older families. The sixth group consists of mideastern industrial cities including Akron, Cincinnati, Buffalo, etc. The cities in this group have high personal income and high residential density, and are oriented toward public transit. The seventh group includes young southwestern areas such as Dallas, Houston, Phoenix, etc. They are characterized by the lowest residential density and public transit orientation. The eighth group consists of Florida cities such as Miami, Tampa, Orlando, etc. These cities have a significant retired population and low residential density. The ninth group contains young northern industrial areas of Davenport, Dayton, Omaha, etc., which have high personal income. Because the nine homogeneous groups were classified according to their arterial transportation needs and requirements, each group (except the first two groups that have only one and two cities, respectively) can be used as an independent dummy variable to account for the spatial variation of geographic characteristics among urban areas.

As indicated previously, little work has been done to incorporate all aforementioned urban characters (except for the urban activity concentration and part of urban size class) into a single urban character relative to travel demand. This research is the first to take such a comprehensive approach for examining the spatial transferability of demand models.

Activity and Transportation System Variables

Six activity system variables most commonly used in the zonal travel demand analysis are population, land area, household size, income, automobile ownership, and activity density. The literature review of other studies has indicated that the relationships of travel demand with these variables are consistently stable, thus they were selected for model formulation (18).

Better transportation systems encourage people to use transportation facilities more often. Thus, the selection of transportation system variables was based on the representativeness of variables to the system's level of service. This study considered the linear feet per capita of the arterial freeway, as well as commuter rail and bus service routes, to reflect transportation system supplies. The supply of roads was estimated by the sum of arterial linear feet per capita plus 2.7 times the number of freeway linear feet per capita (19). On the other hand, the supply of transit was estimated by an assumption that the service efficiency for commuter rail and rapid transit is four times that for bus. Therefore, 1 mile of commuter rail or rapid transit is equivalent to 4 miles of bus service route.

Note that the U.S. Department of Transportation directly or indirectly uses many of the aforementioned activity and transportation system variables as factors for the distribution of federal funds. According to the urban transit formula program of the 1978 Surface Transportation Act, the factors for grant allocation are population, population weighted by a factor of density, commuter rail miles, fixed guideway system route miles, and bus seat miles.

DATA COLLECTION

Despite the massive acquisition of travel data in urban areas during the period 1940-1970, little has been done to develop a consistent set of data in different areas that could be used to define a set of relationships between travel demand and urban character for evaluating urban travel demand models.

The data used in this study were collected from a variety of sources (20-23). Most of them were results of individual area transportation studies and are considered reliable. These data provide socioeconomic, demographic, and travel information for 212 urban areas in the United States including 43 nonmetropolitan areas and 169 metropolitan areas. The body of this information affords the possibility to examine factors determining aggregate travel behavior and the transferability of travel demand models. The 1960s data were used for model calibration, while data collected prior to 1960 and after 1970 were used for model validation. The notations

Table 1. List of variables and notations used.

Notation	Variable	Notation	Variable
$T'_{w,d}$	Logarithmic value of average number of person trips per dwelling unit (work)	H	Average number of persons per dwelling unit
$T'_{n,d}$	Logarithmic value of average number of person trips per dwelling unit (nonwork)	H′	Logarithmic value of H
$T_{w,p}$	Average number of person trips per capita (work)	D_p	Population density (persons per mile^2)
$T_{n,p}$	Average number of person trips per capita (nonwork)	I'_f	Logarithmic value of median family income ($)
$L_{w,t}$	Average work trip duration (min)	V	Supply of transit (linear feet per capita)
$L_{n,t}$	Average nonwork trip duration (min)	K	1 for multinucleated urban areas and 0 for core-concentrated cities
M_a	Percentage of automobile travel	S_1	1 for urban size class group with population of 100 000 to 249 999 and 0, otherwise
M_t	Percentage of transit travel	S_3	1 for urban size class group with population of more than 750 000 and 0, otherwise
P′	Logarithmic value of population (number of persons)	G_3	1 for large northeastern cities and 0, otherwise
B	Land area (miles^2)	G_4	1 for southern cities and 0, otherwise
A'_d	Logarithmic value of the average number of automobiles per dwelling unit	G_5	1 for midwestern cities and 0, otherwise
A_p	Average number of automobiles per capita	G_6	1 for mideastern cities and 0, otherwise
A'_p	Logarithmic value of A_p	G_7	1 for young southwestern cities and 0, otherwise

of variables selected from the model calibration are listed in Table 1. The statistical significance of the data sample is described below.

As indicated previously, there are more data available for metropolitan areas than for nonmetropolitan areas. For the metropolitan area trip frequency model estimation, the sample contains 169 observations that account for 69.5 percent of U.S. metropolitan areas, or 49.2 percent of the U.S. metropolitan population. For the metropolitan area trip duration model estimation, the sample consists of 111 observations that have 46.9 percent of U.S. metropolitan areas, or 43.5 percent of the U.S. metropolitan population. There are only 35 observations available for estimating travel distance equations. However, the size of this sample is still statistically significant (14.4 percent of U.S. metropolitan areas, or 22.4 percent of the U.S. metropolitan population). Some 67 metropolitan areas were selected for mode choice model estimation. This data set accounts for 27.6 percent of U.S. metropolitan areas, or 43.2 percent of the U.S. metropolitan population. For nonmetropolitan areas, the sample contains 43 observations for trip frequency model estimation and 22 observations for trip length modeling.

ANALYSIS METHOD

A number of techniques may be used to examine the impact of urban character on travel demand. Some of these techniques such as automatic interaction detection (24) and contingency table analysis (25) have been widely applied to social science for determining classification schemes. This study selected Gujarati's generalized regression dummy variable approach as an analytical tool for several reasons (26). First, this study not only observes the relationship between urban character and travel demand but also tells how strong the relationship is. Next, the relationship between the two can be verified by data other than those used in the model calibration. Finally, the impact of urban character on the relationship between travel demand and a particular independent variable is shown in a constant term or a variable coefficient, or both. Gujarati's technique is capable of accomplishing these objectives.

The intent of Gujarati's generalized regression dummy variable technique is to portion the sample of nonoverlapping subgroups and to detect whether a subgroup can explain more of the variation in the dependent variable than any other such set of subgroups. By using the stepwise regression procedure it would allow differential intercept and slope for each group of study areas entering the model. Assume the selected model structure includes N explanatory variables and M area characters, the generalized dummy variable equation can be expressed by

$$Y = \alpha_o + \Sigma\alpha_j D_j + \Sigma\beta_i X_i + \Sigma\beta_{ij}(D_j X_{ij}) + U_{ij} \qquad \begin{matrix} i = 1, 2, \ldots, N \\ j = 1, 2, \ldots, M \end{matrix} \qquad (1)$$

where

Y = dependent variable (demand measure),
X_i = independent variables (activity and transportation system characteristics),
D_j = 1 if the observation lies in group j (urban character) or 0 otherwise,
α_o = intercept for all subgroups,
α_j = differential intercept for group j,
β_i = slope coefficient of Y with respect to X_i for all subgroups,
β_{ij} = differential coefficient of Y with respect to $D_j X_{ij}$, and
U = stochastic error term.

Among various forms of the regression model, linear, product, exponential, logarithmic, and combination forms were all used for appropriate travel demand measures. These forms were chosen in order to keep the statistical estimation problem tractable and to account for possible nonlinearities:

Linear	$Y = a + bX$
Product	$Y = aX^b$ or $\ln Y = a + b\ln X$
Exponential	$Y = ab^X$ or $\ln Y = a + bX$
Logarithmic	$Y = a + b\ln X$

where

Y = dependent variable,
X = independent variable, and
a,b = parameters to be estimated.

DETERMINATION OF TRANSFERABLE VARIABLES

Explanatory variables included in trip frequency equations for nonmetropolitan areas are household automobile ownership and household size variables; for metropolitan areas, automobile ownership per capita, household size, and population variables. The transferability of these variables and the estimated models are summarized in Table 2. Comparison of the estimated models for both types of urban areas reveals that trip frequency equations for metropolitan areas are more transferable than their nonmetropolitan counterparts.

The type of urban activity concentration (core-

Table 2. Significance of urban character influencing trip frequency.

Trip Purpose	Explanatory Variable	Coefficient	Size Class	Activity Concentration	Geographic Cluster
Nonmetropolitan area					
Work trips	$\underline{A}'_d$	Intercept	NA		G^*_5
	A'_d	Slope	NA		G^{**}_5, G^*_6
	$\underline{H}'$	Intercept	NA	K*	
	H'	Slope	NA	K**	G^{**}_5
	$\underline{A}'_d$, $\underline{H}'$				
	H'	Slope	NA	K***	
Nonwork trips	$\underline{A}'_d$	Intercept	NA	K***	
	$\underline{H}'$	Slope	NA	K***	
	$\underline{A}'_d$, $\underline{H}'$				
	H'	Slope	NA	K***	
Metropolitan area					
Work trips	$\underline{A}'_p$, $\underline{P}'$				
	$\underline{A}'_p$, $\underline{P}'$, $\underline{H}$				
	H	Slope			G^{**}_7
Nonwork trips	$\underline{A}'_p$, $\underline{P}'$				
	$\underline{A}'_p$, $\underline{P}'$, $\underline{H}$				
	H	Slope			G^{**}_7

Notes: NA = not applicable.
Total explanatory variables included in a model are underlined.
Statistical significance is defined by * at the 10 percent level, ** at the 5 percent level, and *** at the 1 percent level.

concentrated versus multinucleated) appears to be a crucial element in determining the difference of both work and nonwork household trip frequency among nonmetropolitan areas, while its impact on the metropolitan area trip frequency per capita is statistically insignificant. The work household trip frequency in nonmetropolitan areas of geographic cluster 5 (midwestern cities) is significantly different from that of other U.S. nonmetropolitan areas if the household trip frequency is explained either by household automobile ownership or by household size, respectively. On the other hand, the influence of activity concentration on metropolitan area trip frequency per capita is not significant. However, the relationship between trip frequency per capita and household size in metropolitan areas of geographic cluster 7 (young southwestern areas) is significantly different from that of the rest of U.S. metropolitan areas.

Investigating the performance of each explanatory variable reveals that the relationship between household trip frequency and household size is not transferable between certain subgroups of nonmetropolitan areas, i.e., nonmetropolitan areas of geographic cluster 5 versus the rest of U.S. nonmetropolitan areas and the core-concentrated versus multinucleated nonmetropolitan areas. For metropolitan areas, the trip frequency model with household size as the independent variable is not transferable between metropolitan areas of geographic cluster 7 and the rest of U.S. metropolitan areas. The relationship between trip frequency per capita and automobile ownership per capita is spatially transferable for metropolitan areas. Such a transferability is applicable to both trip purposes. However, the spatial transferability of the nonmetropolitan area household trip frequency model with household automobile ownership as the independent variable is only limited to work trips.

Table 3 suggests that the spatial transferability of mode choice model is fairly high. Both percentage of travel by transit and automobile occupancy equations are independent of the influence of geographic factors, while the percentage of travel by automobile and the percentage of walking equations are influenced only by the urban character of geographic cluster 3 (large northeastern cities) and geographic cluster 7 (young southwestern cities), respectively. The most important urban size class variable is urban size class 3 (cities with a population of more than 750 000), which indicates the divergency of mode choice models between metropolitan areas with a population of more than 750 000 and metropolitan areas of 750 000 people or less. The comparison between explanatory variables shows that automobile ownership per capita is a transferable variable whereas the transferability of population density, the supply of public transit service, and the income variables are limited to a certain degree.

By using population, land area, and population density as variables to explain trip length, the transferability of the model is relatively low when compared with that of trip frequency and mode choice models. As indicated in Table 4, the slope of the population variable in trip duration for both trip purposes and for both types of urban areas is significantly affected by the type of urban activity concentration. The slope of the population density variable differs among urban area groups as defined by urban size class, while the slope of the land area variable is significantly different between metropolitan areas of geographic cluster 4 (southern cities) and the rest of U.S. metropolitan areas. Because the transferability of explanatory variables is low, the estimated trip length models have less transferability than trip frequency and mode choice models.

In general, the results of the above analyses indicate that the transferability of models depends on the type of demand measure and explanatory variable as well as the number of explanatory variables included in the model. The combination of two or more explanatory variables in a model tends to enhance transferability.

MODEL VERIFICATION

Verification of the findings presented here can be accomplished by the traditional case-study approach or the validation of the developed models on which the preceding analysis was based. However, in order to exhaustively examine the validity of the findings the traditional approach would require 4x30 (120) case studies for nonmetropolitan areas and 6x90 (540) case studies for metropolitan areas. Thus the second method of verification was selected in this study. The developed models were applied to real-life situations. The validity of these models was evaluated by comparing the estimated and actual travel demand. The results of applications are presented below. Note that the selection of study areas was based on data availability, and the data for use in validation were collected either before 1960 or after 1970.

Demand Models for Nonmetropolitan Areas

Trip Frequency Models

The trip frequency equations contained in Table 5 were applied to five U.S. nonmetropolitan areas to forecast the total person trips per dwelling unit. The forecast years (forward and backward) range from 1947 to 1948 and from 1972 to 1973. The areas, which were selected in accordance with the diversity of urban character, include Nashaua, New Hampshire; Tri-cities, Virginia; Rapid City, South Dakota; Tucson, Arizona; and Pontiac, Michigan. The prediction error ranges from 2 to 13.7 percent. The result of this application indicates that the predictive ability of trip frequency models is satisfactory.

Trip Length Models

The equations for applying trip length models to

forecast the average trip duration for U.S. nonmetropolitan areas are shown in Table 5. The study areas are Tri-Cities; Beloit, Wisconsin; Hutchinson, Kansas; and Tallahassee, Florida. The forecast years range from 1971 to 1973. Comparison of the estimated and actual trip durations shows that the forecast error falls into a range of 3.1-10.4 percent. The accuracy of these estimations is considered acceptable.

Table 3. Significance of urban character influencing mode choice and automobile occupancy for metropolitan areas.

Transportation Mode	Explanatory Variable	Coefficient	Size Class	Activity Concentration	Geographic Cluster
Percentage of automobile travel	$\underline{A_p}$, $\underline{D_p}$ / D_p	Slope			G_3^{**}
Percentage of transit travel	$\underline{A_p}$, $\underline{D_p}$, $\underline{P'}$ / D_p	Slope	S_3^{**}		
	P'	Slope	S_1^{**}		
Percentage of walk	$\underline{A_p}$, $\underline{V}$ / V	Slope	S_3^{**}		G_7^{**}
Automobile occupancy	$\underline{I_f'}$	Slope	S_3^{*}		

Notes: Total explanatory variables included in a model are underlined.
Statistical significance is defined by * at the 10 percent level, ** at the 5 percent level, and *** at the 1 percent level.

As to mode choice models, nonmetropolitan areas generally have a limited choice of travel modes. The data limitation and necessity precluded verification of mode choice models for nonmetropolitan areas.

Demand Models for Metropolitan Areas

Trip Frequency Models

Two trip frequency equations as shown in Table 6 were applied to forecast the average work and nonwork person trips per capita for seven U.S. metropolitan areas: Phoenix, Arizona; Wichita, Kansas; Huntington, West Virginia; Akron, Ohio; Anderson, Indiana; Denver, Colorado; and Duluth-Superior, Minnesota. The population of these metropolitan areas ranges from 90 000 (Anderson) to 1 220 000 (Phoenix), while the forecast years range from 1970 to 1976. The result indicates that the predictive ability of work trip frequency model is satisfactory, with an error ranging from 0 to 9.3 percent. The prediction error for the nonwork person trip frequency equation is varied. It ranges from 0.3 percent in Wichita to 22.8 percent in Denver, and to 39.5 percent in Phoenix. The main cause of the large discrepancy between the estimated and actual is the dramatic change of automobile ownership. For example, the automobile ownership for Phoenix was

Table 4. Significance of urban character influencing trip length.

Trip Purpose	Explanatory Variable	Coefficient	Size Class	Activity Concentration	Geographic Cluster
Nonmetropolitan area					
Work trips					
Trip duration	$\underline{P'}$	Slope	NA	K**	G_5^{*}, G_6^{**}
Travel distance	$\underline{P'}$	Slope	NA		G_6^{**}
Nonwork trips					
Trip duration	$\underline{P'}$	Slope	NA	K*	G_5^{**}, G_6^{***}
Travel distance	$\underline{P'}$	Slope	NA		G_6^{**}
Metropolitan area					
Work trips					
Trip duration	$\underline{P'}$, $\underline{B}$, $\underline{D_p}$				
	P'	Slope		K***	G_7^{**}
	B	Slope			G_4^{**}
	D_p	Slope	S_1^{**}, S_3^{**}		
Travel distance	$\underline{P'}$, $\underline{B}$				
	$\overline{B}$	Slope			G_4^{*}
Nonwork trips					
Trip duration	$\underline{P'}$, $\underline{B}$, $\underline{D_p}$	Intercept			G_3^{*}
	P'	Slope		K***	
	B	Slope			G_4^{**}
	D_p	Slope	S_1^{**}	K**	
Travel distance	$\underline{P'}$, $\underline{B}$				
	$\overline{P'}$	Slope	S_1^{***}		
	B	Slope	S_2^{**}	K***	G_4^{**}

Notes: NA = not applicable.
Total explanatory variables included in a model are underlined.
Statistical significance is defined by * at the 10 percent level, ** at the 5 percent level, and *** at the 1 percent level.

Table 5. Trip frequency and trip length models for nonmetropolitan areas.

Factor	Equation	
Work trip frequency	$T'_{w,d} = -0.702 + 1.026\ A'_d + 1.285\ H' - 0.105\ (KH')$ [29.0]*** [33.3]*** [4.3]***	$R^2 = 0.8013$ D.F. = 28
Nonwork trip frequency	$T'_{n,d} = 1.835 + 1.080\ A'_d - 0.251\ (KH')$ [10.2]*** [6.6]***	$R^2 = 0.3728$ D.F. = 29
Work trip duration	$L_{w,t} = -30.671 + 3.506\ P' + 0.181\ (G_5P') + 0.232\ (G_6P') - 0.168\ (kP')$ [34.3]*** [2.1]* [8.7]*** [4.2]**	$R^2 = 0.7821$ D.F. = 25
Nonwork trip duration	$L_{n,t} = -26.180 + 2.896\ P' - 0.131\ (KP') + 0.139\ (G_5P') + 0.248\ (G_6P')$ [27.2]*** [3.0]* [7.6]*** [11.6]***	$R^2 = 0.7618$ D.F. = 25

Notes: [ ] = F-value.
* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

Table 6. Trip frequency, trip length, and mode choice models for metropolitan areas.

Factor	Equation
Work trip frequency	$T_{w,p} = 0.708 + 0.939\,A_p - 0.0316\,P' - 0.0143\,(G_7H)$ [24.7]*** [16.6]*** [3.0]*** $R^2 = 0.2915$ D.F. = 110
Nonwork trip frequency	$T_{n,p} = 3.001 + 4.300\,A_p - 0.193\,P' - 0.056\,(G_7H)$ [37.9]*** [41.1]*** [3.0]*** $R^2 = 0.4121$ D.F. = 110
Work trip duration	$L_{w,t} = -17.288 + 2.472\,P' - 0.183\,(KP') - 0.153\,(G_7P') + 0.000\,331\,B + 0.0016\,(G_4B) - 0.0086\,(S_1D_p) + 0.0038\,(S_3D_p)$ [11.7]***[9.8]*** [3.2]*** [3.4]*** [13.4]*** [6.1]*** [7.9]*** $R^2 = 0.7702$ D.F. = 103
Nonwork trip duration	$L_{n,t} = -4.729 + 2.268\,G_3 + 1.225\,P' - 0.0227\,(KP') + 0.0012\,(G_4B) + 0.0019\,(KD_p) - 0.0087\,(S_1D_p)$ [2.6]** [5.4]*** [8.8]*** [8.9]*** [2.4]** [8.9]*** $R^2 = 0.5729$ D.F. = 104
Percentage of automobile travel	$M_a = 45.536 + 95.773\,A_p - 0.0044\,(G_3D_p)$ [47.3]*** [9.3]*** $R^2 = 0.4821$ D.F. = 64
Percentage of transit travel	$M_t = 31.143 - 64.377\,A_p + 0.491\,(S_3D_p') - 0.247\,(S_1P')$ [11.4]*** [12.8]*** [2.2]** $R^2 = 0.5795$ D.F. = 63

Notes: [ ] = F-value.
* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

near 0.6 cars per capita for the forecasting year, while the average automobile ownership of the 1960s sample was 0.37 cars per person with a range of 0.25-0.45 cars per capita. The forecast error for total person trips falls into a range of 0.4-31.9 percent. The accuracy of forecast is still comparable to that of zonal models.

Trip Length Models

Six U.S. metropolitan areas were selected to evaluate the predictive ability of trip length models for metropolitan areas: Phoenix; Lima, Ohio; Wichita; Akron; Denver; and Duluth-Superior. Two models for this evaluation are contained in Table 6. The forecast years range from 1970 to 1976. Comparison between the actual and estimated shows that both models perform fairly well. The discrepancy between them ranges from 2.2 to 10.0 percent for the work trip duration equation and from 2.0 to 12.0 percent for the nonwork trip duration model. The results indicate that the trip duration equations are capable of estimating trip duration for various metropolitan areas with a diversity of demographic and geographic characteristics.

Mode Choice Models

The developed mode choice models as shown in Table 6 were applied to forecast the modal-split between automobile driver, automobile passenger, and public transit for three U.S. metropolitan areas of Akron, Denver, and Duluth-Superior. The forecast year of the three applications is 1970. Comparison of the estimated and actual percent of each mode reveals that the developed models can predict the mode choice of the three metropolitan areas with a high degree of accuracy. The prediction errors are less than 5 percent.

The results of these model verifications indicate that the predictive ability of the models developed in this study is comparable to specific models developed for particular areas as shown in zonal experience. They also imply that the relationships between travel demand and urban character as identified in this study are stable over time.

CONCLUSIONS

The purpose of this paper has been to call attention to the consideration of certain types of urban character when the transfer of travel demand models is desired. A generalized regression dummy variable approach was used to identify the effects of urban character on travel demand. The results provide some guidelines for determining model transferability.

The effects of urban character on travel demand depend on the aspect of demand measure and the model specification. In accordance with the selected variables this study has found no models (including trip frequency models) that are perfectly transferable. However, with careful selection, the transfer of a model from one geographic areas to another is not impossible.

The reliability of these findings was verified by testing the models that were developed for examining the relationships between urban character and travel demand measures. These models were applied to several U.S. urban areas with satisfactory results. The findings of this study provide the following conclusions.

1. For nonmetropolitan area work trips, if the household trip frequency is explained by household automobile ownership, the model is not transferable between three urban classes: midwestern cities, mideastern cities, and the rest of U.S. cities. If the household trip frequency is explained by household size, both activity concentration (core-concentrated versus multinucleated) and geographic cluster (midwestern cities versus the rest of U.S. cities) become the dominant factors for determining model transferability. If both household automobile ownership and household size are included in the model, the model is not transferable between cities with different types of activity concentration.
2. For nonmetropolitan areas, the nonwork trip frequency model composed of household automobile ownership and/or household size variables is significantly affected by the type of activity concentration.
3. For metropolitan areas, both work and nonwork trip frequency models composed of automobile ownership per capita and household size are not transferable between young southwestern cities and the rest of U.S. cities.
4. For metropolitan area mode choice, if the percentage of automobile travel is explained by both automobile ownership per capita and population density, the model is not transferable between large northern cities and the rest of U.S. metropolitan areas. If the percentage of transit travel model is composed of automobile ownership per capita, population density, and population, the model is not transferable among cities with population 50 000 to

99 999; 100 000 to 249 999; 250 000 to 750 000; and more than 750 000. If the percentage of walk is explained by both automobile ownership per capita and transit system supply, the model is not transferable between cities with a population equal to, less than, or more than 750 000 and between young southwestern cities and the rest of U.S. cities. If automobile occupancy is explained by the median family income the urban size becomes a significant dummy variable.

5. For nonmetropolitan areas, the trip length is explained by population and the model is not transferable between different types of activity concentration and between geographic clusters of midwestern cities, mideastern cities, and the rest of U.S. cities for trip duration, nor between mideastern cities and the rest of U.S. cities for travel distance.

6. For metropolitan areas, the trip length is explained by population, land area, and population density. The model transferability is limited to urban groups defined by activity concentration, geographic cluster, and urban size class.

REFERENCES

1. H.J. Nelson. A Service Classification of American Cities. Economic Geography, Vol. 31, 1955, pp. 189-210.
2. V. Jones and A. Collver. Economic Classification of Cities and Metropolitan Areas. In Municipal Year Book, International City Management Assn., Washington, DC, 1960, pp. 67-79, 89-90.
3. R.L. Forstall. Economic Classification of Places Over 10 000: 1960-1963. In Municipal Year Book, International City Management Assn., Washington, DC, 1967, pp. 30-65.
4. W.H. Botting and B.T. Goley. A Classification of Urbanized Areas for Transportation Analysis. HRB, Highway Research Record 194, 1967, pp. 32-61.
5. A.M. Voorhees and S.J. Bellomo. Urban Travel and City Structure. HRB, Highway Research Record 322, 1970, pp. 121-135.
6. T.F. Golob, E.T. Canty, and R.L. Gustafson. Classification of Metropolitan Areas for the Study of New Systems of Arterial Transportation. General Motors Research Laboratories, Warren, MI, Aug. 1972.
7. B.J.L. Berry and K.B. Smith (eds.). City Classification Handbook: Methods and Applications. Wiley-Interscience, New York, 1972.
8. Projection of Urban Person Transportation Demand. Peat, Marwick, Livingston, and Co., New York, March 1968.
9. L.C. Caldwell III and M.J. Demetsky. Transferability of Trip Generation Models. TRB, Transportation Research Record 751, 1980, pp. 56-62.
10. R.J. Williams and P.H. Wright. Factors Which Influence Modal Choice. Traffic Quarterly, Vol. 33, No. 1, Jan. 1974, pp. 271-289.
11. M. Williams. Factors Affecting Mode Choice Decisions in Urban Travel. Transportation Research, Vol. 12, 1978, pp. 91-96.
12. F-L Ou and J.C. Yu. Urban Characteristics Related to Travel Demand Model Transferability. In Proc., 11th Annual Modeling and Simulation Conference, Pittsburgh, May 1980, pp. 1097-1101.
13. B.S. Pushkarev and J.M. Zupan. Public Transportation and Land Use Policy. Indiana Univ. Press, Bloomington, 1977.
14. W.J. Fogarty. Trip Production Forecasting Models for Urban Areas. Transportation Engineering Journal of ASCE, Vol. 102, No. TE4, Nov. 1976, pp. 831-845.
15. J.W. Clark. Assessing the Relationship Between Urban Form and Travel Requirements. Urban Transportation Program, Univ. of Washington, Seattle, Res. Rept. 75-6, 1975.
16. Y. Chan and F-L Ou. Tabulating Demand Elasticities for Urban Travel Forecasting. TRB, Transportation Research Record 673, 1978, pp. 40-46.
17. H.P. Friedman and J. Rubin. On Some Invariate Criteria for Grouping Data. Journal of the American Statistical Assn., Vol. 62, 1967, pp. 1159-1178.
18. F-L Ou. Generalization of Demand Models for Areawide Travel Forecasting. Department of Civil Engineering, Univ. of Utah, Salt Lake City, Ph.D. dissertation, July 1980.
19. F.S. Koppelman. Preliminary Study for Development of a Macro Urban Travel Demand Model. U.S. Department of Transportation, 1972.
20. Alan M. Voorhees and Assoc., Inc. A System Sensitive Approach for Forecasting Urbanized Area Travel Demand. U.S. Department of Transportation, Dec. 1971.
21. F.B. Curran and J.T. Stegmaier. Travel Patterns in 50 Cities. HRB, Bull. 203, 1958, pp. 99-130.
22. Journal to Work, Population Census. Bureau of the Census, U.S. Department of Interior, 1960 and 1970.
23. Y. Chan, F-L Ou, J. Perl, and E. Regan. Review and Compilation of Demand Forecasting Experiences: An Aggregation of Estimation Procedure. Office of Univ. Research, Office of the Secretary, U.S. Department of Transportation, June 1977.
24. J.A. Sonquist. Multivariate Model Building. Univ. of Michigan, Ann Arbor, 1970.
25. M. Rosenberg. The Logic of Survey Analysis. Basic Books, New York, 1968.
26. D. Gujarati. Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Linear Regression: A Generalization. American Statistics, Vol. 24, No. 5, Dec. 1970, pp. 18-22.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Goodness-of-Fit Measures and the Predictive Power of Discrete Choice Models

CARLOS F. DAGANZO

Although a number of goodness-of-fit measures for discrete choice models have been proposed and are widely in use, there have been few attempts at interpreting their physical meaning at a practical level. This paper presents a family of goodness-of-fit measures, which contains currently used measures such as the pseudo-correlation coefficient and the percent right, and shows how its members are related. More important, it is shown that one of these measures has an interpretation identical to the correlation coefficient of multiple regression in that it can be used to calculate the expected reduction in the root-mean-square prediction error afforded by a model. This goodness-of-fit measure, d, is uniquely defined for binary models, and can be approximated by an easy to calculate consistent statistic, D. For models with more than two alternatives, the same can be done, but the measure depends on the alternative under consideration. The d-measure usually takes values in between the commonly used normalized percent right measure and the pseudo-correlation coefficient.

A discrete choice model is a formula that relates some explanatory variables, X (sometimes called attributes), to the probability that an individual chooses one of a set of alternatives. The expression given below is the discrete choice model itself. It assumes that there are I alternatives, i, and is called the choice function:

$$\Pr[\text{choose } i|X] = P_i(X),\ i = 0, 1, \ldots, I-1 \tag{1}$$

The closer the probabilities given by Equation 1 are to either zero or one, the more unequivocal will be the predictions of the model. A true model--it is assumed that Equation 1 does not contain specification errors--that gives values close to either zero or one for most of the values of X that are likely to be encountered in practice is superior to a model that gives probabilities that barely depend on X.

The distinction of good and bad models can be captured by many measures of goodness (of fit), and a purpose of this paper is to discuss these in a unified way. The multiple linear regression analog of these measures is the correlation coefficient. Ultimately, because a model is as good as its predictions, it will be shown that there is a goodness-of-fit measure, d, which is related to prediction error reduction in the same way as the correlation coefficient of regression analysis. Thus, the model goodness can be readily interpreted by analysts used to working with regression models.

The d-measure will be related to other commonly used goodness-of-fit indicators such as the "percent predicted right" indicator of success tables and the pseudo-correlation coefficient of McFadden (1). Hauser (2) provides a good review of goodness-of-fit measures to which the reader is referred.

ON MS PREDICTION ERRORS FOR REGRESSION MODELS

The linear regression model is

$$Y = \beta \cdot X^T + \epsilon \tag{2}$$

where Y is a continuous dependent variable, β and X are row vectors of constants and explanatory variables, and ε is a zero mean random error term that is independent of X and of the error terms of other observations.

The correlation coefficient, ρ, of a regression model (see Equation 2) can be expressed as follows:

$$1 - \rho^2 = [\text{var}(\epsilon)]/[\text{var}(Y)] \tag{3}$$

This is well known, and we also note that

$$\begin{aligned}\text{var}(Y) &= E_X[\text{var}(Y|X)] + \text{var}_X[E(Y|X)] \\ &= [\text{var}(\epsilon)] + \text{var}\,[(\beta \cdot X^T)]\end{aligned} \tag{4}$$

In the regression terminology, var(ε) is called the unexplained variance and var(β • X^T) the explained variance. It should be noted that the ratio of explained to total variance (ρ^2) can be increased for the same real world phenomena by increasing the spread in the sampling distribution of X. It cannot be increased, however, by increasing the sample size.

The correlation coefficient can also be related to an average reduction in prediction error that is afforded by the model if the sampling distribution of X coincides with the distribution of values for which predictions are desired.

For a model without explanatory variables, the predictions will be constant, y_o. The mean squared prediction error for a model without explanatory variables, MSE, when the prediction is y_o is

$$MSE = E(Y - y_o)^2 = [\text{var}(Y)] + [E(Y) - y_o]^2$$

Because E(Y) is the value of y_o that minimizes the MSE, we assume that $y_o = E(Y)$ and denote it by $\bar{Y}$. Then,

$$MSE = \text{var}(Y) \tag{5}$$

We note that the MSE for any given value of X, MSE_X, is

$$\begin{aligned}MSE_X &= E(Y - \bar{Y}^2|X) \\ &= \text{var}(Y - \bar{Y}|X) + (\beta \cdot X^T - \bar{Y})^2 \\ &= \text{var}(Y|X) + [\beta \cdot (X - \bar{X})^T]^2 \\ &= \text{var}(\epsilon) + [\beta \cdot (X - \bar{X}^T)]^2\end{aligned} \tag{6}$$

That is, MSE_X is made up of a random component, var(ε), and a systematic component, $[\beta \cdot (X - \bar{X})^T]^2$. As shown below, the systematic component is removed by the model with explanatory variables. Because Equation 5 is the average of Equation 6 over the distribution of X: $MSE = E_X(MSE_X)$. MSE will be called the average mean squared prediction error.

For the model with explanatory variables, the mean squared prediction error for a given value of X, MSE_X^W, can be shown to be independent of X. Since the model prediction is β • X^T, we may write

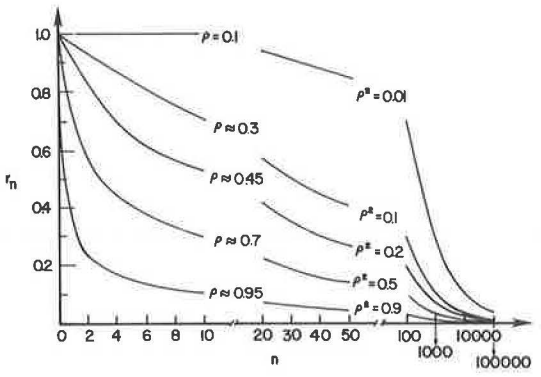
$$MSE_X^W = E[(Y - \beta \cdot X^T)^2|X] = E[\epsilon^2|X] = \text{var}(\epsilon) \tag{7}$$

Thus, the average mean squared prediction error, MSE^W, is also given by var(ε). Clearly, then, the interpretation of the correlation coefficient, ρ, that is related to prediction errors is

$1 - \rho^2$ = average mean squared prediction error with model ÷ average mean squared prediction error without model

and one can interpret $\sqrt{1 - \rho^2}$ as the after/

Figure 1. Residual prediction error as a fraction, r_n, of the prediction error without a model for different values of the correlation coefficient, ρ, and the prediction group size, n.



before ratio of prediction errors that can be expected. This ratio will be noted by r.

Instead of predicting the dependent variable for one observation, if one wants the average for a set of n observations with the same value of X, the prediction errors will be reduced. A bar will be used over the MSE symbol to denote that Y is an average of n observations. The same arguments, as before, yields

$$\overline{MSE^W} = \overline{MSE_X^W} = (1/n)\text{var}(\epsilon) \tag{8}$$

This is the equivalent of Equation 7. For the model without explanatory variables, only the random component of the prediction errors is diminished:

$$\overline{MSE_x} = (1/n)\text{var}(\epsilon) + [\beta \cdot (X - \bar{X})^T]^2 \tag{9}$$

Of course,

$$\overline{MSE} = E_x(\overline{MSE_x}) = (1/n)\text{var}(\epsilon) + \text{var}(\beta \cdot X^T) \tag{10}$$

which is the equivalent of Equation 4. The before/after prediction error ratio for an average of n observations is defined as above:

$$r_n = \sqrt{MSE^W/\overline{MSE}} = \sqrt{(1-\rho^2)/(1-\rho^2) + n\rho^2} \tag{11}$$

This well-known expression is plotted on Figure 1 to clarify further the relationship between ρ and prediction errors.

The next section contains a similar discussion for binary discrete choice models and introduces a measure, d, of the difference in prediction accuracy with and without a model. The rest of the paper attempts to relate the different goodness-of-fit measures that exist.

THE d-MEASURE

Let us consider binary discrete choice models. For these models $P_1(X) = 1 - P_0(X)$, and the dependent variable is either zero or one. It is assumed, however, that $P_i(X)$ is strictly between zero and one. For any given value of X, the expected value of the Bernoulli dependent variable, i, is $P_1(X)$. With this in mind, we can repeat the steps noted in the earlier section on MS prediction errors for regression models.

If y_o is the predicted value of i for a model without explanatory variables, the MSE is

$$MSE = E(i - y_o)^2 = \text{var}(i) + [E(i) - y_o]^2$$

The value of y_o that minimizes MSE is E(i), the fraction of the population of individuals for which predictions are desired that chooses alternative 1. This value will be called f_1 and clearly $\text{var}(i) = f_1(1 - f_1)$. The trivial cases with $f_1 = 0$ or $f_1 = 1$ will not be considered in this paper. This is consistent with the postulate that $P_1(X)$ is different from zero and one since there are no specification errors. From now on, we assume that $y_o = f_1$. Thus,

$$MSE = f_1(1 - f_1) \tag{12}$$

which is the equivalent of Equation 5.

Analogously,

$$MSE_x = E[(i - f_1)^2|X] = \text{var}(i|X) + [E(i|X) - f_1]^2 = P_1(X)[1 - P_1(X)] + [P_1(X) - f_1]^2 \tag{13}$$

As in Equation 6, $P_1(X)[1 - P_1(X)]$ is a random component and $[P_1(X) - f_1]^2$ is a removable systematic component. The difference is that the random component is not fixed. Nevertheless, it is true that $MSE = E_X(MSE_X)$, and MSE can still be called legitimately the average mean squared prediction error.

For models with explanatory variables, MSE_x^W is no longer independent of X. Let $y_o(X)$ be the prediction, and write

$$MSE_x^W = E\{[i - y_o(X)]^2|X\} = \text{var}(i|X) + [E(i|X) - y_o(X)]^2 = P_1(X)[1 - P_1(X)] + [P_1(X) - y_o(X)]^2$$

Since a prediction of $y_o(X) = P_1(X)$ minimizes the mean squared error, it will be assumed to be that way from now on. Thus,

$$MSE_x^W = P_1(X)[1 - P_1(X)] \tag{14}$$

which as with regression is the random component of MSE_X.

Unlike in regression, however, the average mean squared prediction error must be calculated. It is

$$MSE^W = E_X(MSE_x^W) = E_X[\text{Var}(i|X)] = E\{P_1(X)[1 - P_1(X)]\} \tag{15}$$

The mean of the (removed) systematic component is

$$E_X\{[P_1(X) - f_1]^2\} = \text{var}_X[P_1(X)]$$

which is similar to the expression of the systematic error of the regression model.

By analogy to the correlation coefficient of regression, let us define a measure of goodness-of-fit, d, which will capture the difference that the model makes:

$$1 - d^2 = MSE^W/MSE = E_X\{P_1(X)[1 - P_1(X)]\}/f_1(1 - f_1) \tag{16a}$$

Alternatively,

$$d^2 = \text{var}_X[P_1(X)]/f_1(1 - f_1) \tag{16b}$$

It is not difficult to realize that Figure 1 also applies to the d-measure. This interpretation is even more important for discrete choice models because (at least in most transportation applications) one is usually interested in predicting the behavior of groups of people with more than one person ($n > 1$).

ESTIMATION OF d: THE D-STATISTIC

The value of d can be approximated from the data used for estimation. Assume that a random, attribute-

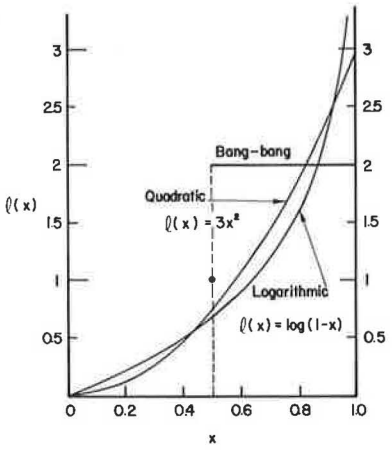


Figure 2. Quadratic, logarithmic, and bang-bang loss functions.

based sample is used and define $\bar{P}_1$ equals the mean estimated choice probability of alternative 1 in the sample, and S_p^2 equals the sample variance of the estimated choice probability for alternative 1 (or 0--they are the same) in the sample.

Then, the following statistic,

$$D = \sqrt{S_p^2/\bar{P}_1(1-\bar{P}_1)} \tag{17}$$

intuitively approximates d for large samples. Appendix A proves that under regularity conditions, D is consistent.

To show that $D \in [0, 1]$, consider the set of estimated alternative 1 choice probabilities for all the observations in the sample. The sample mean of these probabilities was denoted $\bar{P}_1$ and the second sample moment will be denoted $\bar{\bar{P}}_1$. Because all the probabilities are between zero and one, we may write $\bar{\bar{P}}_1 \leq \bar{P}_1$, or subtracting $\bar{P}_1^2$ on both sides, $\bar{P}_1 - \bar{P}_1^2 \leq \bar{P}_1(1 - \bar{P}_1)$. Since the sides of this inequality are the numerator and denominator of Equation 17, it is clear that $D \leq 1$. That $D \geq 0$ is obvious.

FRAMEWORK FOR COMPARISON

Let us now investigate how the d-measure is related to other measures currently in use. To do this more easily, the measures will be derived from a common model.

Let us assume that when an observation is taken, one experiences the following penalty for predicting y as the choice probability for alternative 1 [$y \in (0, 1)$] when the choice is actually i (i = 0, 1):

$$L(y, i) = \ell(|y - i|),$$

where $\ell(\cdot)$ is a real-valued, nondecreasing loss function defined in the (0, 1) interval, with $\lim_{x \to 0} \ell(x) = 0$. The following are three possibilities for $\ell(x)$; they are labeled A, B, and C since they will be used in the sequel:

(A) Quadratic loss: $\ell(x) = x^2$.

(B) Logarithmic loss: $\ell(x) = -\log(1 - x)$

(C) Bang-bang loss:
$$\ell(x) = \begin{cases} 0 & \text{if } x < (1/2) \\ 1/2 & \text{if } x = (1/2) \\ 1 & \text{if } x > (1/2). \end{cases}$$

Figure 2 depicts these functions. (Because the scaling of the loss functions does not affect the goodness-of-fit measures that one derives, the scales in the figure are such that $\int_0^1 \ell(x)dx = 1$ in all three cases.)

The goodness of a model can, therefore, be measured by comparing the expected loss for a prediction with and without the model.

If we do not have a model, a prediction, $y = P_1$, that minimizes the expected loss (if it exists) is found by

$$\min_{y \in (0,1)} \left\{E_i[\ell(|y - i|)]\right\} = \min_{y \in (0,1)} \left\{f_1 \ell(1 - y) + (1 - f_1)\ell(y)\right\} \tag{18}$$

where, as before, f_1 is the fraction of the population choosing 1. As is commonly done in Bayesian analysis, only loss functions for which $P_1 = f_1$ are considered. Otherwise, there would be a reward for providing the wrong prediction, and this should be avoided. Loss functions for which $P_1 = f_1$ will be called regular.

For regular loss functions that are differentiable in [0, (1/2)], the derivative, $\ell'(x)$, of the loss function must satisfy the following (necessary) symmetry condition in (0, 1) if the minimum of Equation 18 is to be equal to f_1:

$$[\ell'(x)]/x = \ell'(1 - x)/1 - x, x \neq (1/2) \tag{19}$$

Cases A, B, and C all satisfy Equation 19 and are regular.

The minimum of Equation 18 is unique only for cases A and B, however.

A regular loss function that is nondecreasing and differentiable in [0, (1/2)] may have a discontinuity at x = (1/2). (See case C.)

For regular loss functions, the expected loss without a model, L, is calculated with a prediction, $P_1 = f_1$:

$$L = \psi(f_1) = f_1 \ell(1 - f_1) + (1 - f_1)\ell(f_1) \tag{20}$$

The next theorem establishes some important properties of $\psi(x)$.

Theorem 1: Assume $\ell(x)$ is real-valued, nonnegative, nondecreasing, regular, differentiable in [0, (1/2)], and such that $\lim_{x \to 0} \ell(x) = 0$. Then, $\psi(x)$ is continuous, bounded, concave, reaches a maximum at x = 1/2, and is such that $\lim_{x \to 0,1} \psi(x) = 0$.

Proof: Because it is nonnegative, $\psi(x)$ is bounded and must satisfy

$$\psi(x) \leq x\ell(1/2) + (1 - x)\ell(1/2) = \ell(1/2)$$

since by regularity $\psi(x)$ is the g.l.b. of Equation 18. Thus, $\ell(1/2)$ bounds $\psi(x)$ if $\psi(x)$ is regular. Since $\psi(1/2) = \ell(1/2)$, $\psi(x)$ reaches a maximum at x = 1/2.

To prove that $\psi(x)$ is continuous, it suffices to show that $\psi(x) \to \psi(1/2)$ as $x \to (1/2)$.

Because $\ell(x)$ is regular and $\psi(1/2) = \ell(1/2)$, Equation 18 with $f_1 = (1/2)$ enables us to write:

$$\psi(1/2) = \ell(1/2) \leq (1/2)\left\{\ell[(1/2) + \epsilon] + \ell[(1/2) - \epsilon]\right\}, \forall \epsilon \in [0,(1/2)] \tag{21}$$

For $f_1 = (1/2) + \epsilon$, Equation 18 yields

$$\psi[(1/2) + \epsilon] = 1/2\left\{\ell[(1/2)+\epsilon] + \ell[(1/2) - \epsilon]\right\} - \epsilon\left\{\ell[(1/2) + \epsilon] - \ell[(1/2) - \epsilon]\right\} \leq \psi(1/2), \forall \epsilon \in [0, (1/2)] \tag{22}$$

The inequality stems from the fact that $\psi(1/2)$ is a maximum of $\psi(x)$. It is clear from Equations 21 and 22 that

$$|\psi[(1/2)] - \psi[(1/2) + \epsilon]| \leq \epsilon\left\{\ell[(1/2)+\epsilon] - \ell[(1/2) - \epsilon]\right\}, \forall \epsilon \in [0,(1/2)]$$

and since the RHS goes to zero when $\epsilon \to 0$, $\psi[(1/2) + \epsilon] \to \psi[(1/2)]$. This establishes continuity. Concavity is next.

By using Equation 19 we see that $\psi'(x) = \ell(1 - x) - \ell(x)$, $x \in [0, (1/2)]$. Clearly, $\psi'(x)$ is nonnega-

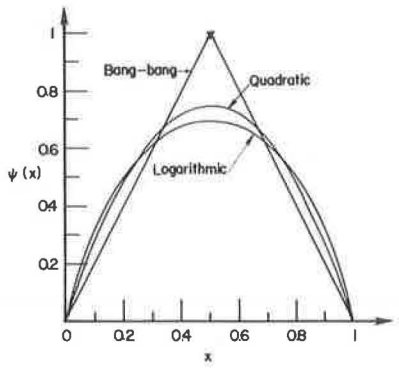


Figure 3. $\psi(x)$ for quadratic, logarithmic, and bang-bang loss functions.

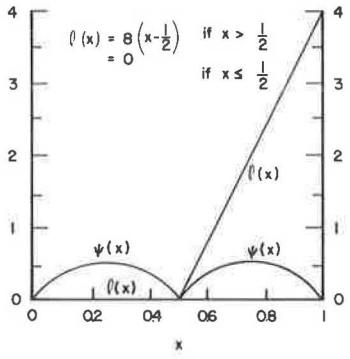


Figure 4. Example of irregular loss function.

tive and nonincreasing. Thus, $\psi(x)$ is concave and nondecreasing in [0, (1/2)].

Since $\psi(x)$ is continuous at 1/2 it is also continuous, concave, and nondecreasing in [0, (1/2)]. The concavity of $\psi(x)$ in (0, 1) follows immediately from the above properties and the symmetry of $\psi(x)$ around 1/2. (Alternatively, the reader can verify that $\{\partial\psi(x) = \psi'(x)$ if $x \neq (1/2)$, $\partial\psi(x) = 0$ if $x = (1/2)\}$ is a subgradient of $\psi(x)$.

To show that $\lim_{x\to 0,1} \psi(x) = 0$, it suffices to show that $(1 - x)\ell(x) \to 0$ as $x \to 1^-$. If $\ell(x)$ is bounded in a neighborhood of 1, this is obvious. Otherwise, $\ell(x) \to \infty$ as $x \to 1^-$ and the limit is the same as

$$\lim_{x\to 1^-} (1 - x)\ell(x) = \lim_{x\to 1^-} (1 - x)^2\ell'(x)$$

because of l'Hôpital's rule. ($[\ell(x)]$ is differentiable in [(1/2), 1] because of the symmetry condition implied by regularity.) By using Equation 19,

$$\lim_{x\to 1^-} (1 - x)^2\ell'(x) = \lim_{x\to 1^-} x(1 - x)\ell'(1 - x) = 0 \cdot \text{QED}$$

Figure 3 plots $\psi(x)$ for cases A, B, and C. Note in particular that $\psi(x)$ is continuous at x = (1/2) for the bang-bang case. Figure 4 illustrates that the theorem does not hold for loss functions that are not regular.

The next step will be to calculate the expected prediction loss when we use the true model $P_1(X)$ and to compare that with $\psi(f_1)$ to derive a goodness-of-fit measure that will capture the reduction in loss achieved by the model.

Goodness-of-Fit Measures from Regular Loss Functions

Given X, the probability that i = 1 is $P_1(X)$ and because the loss function is regular, the optimal prediction is $P_1(X)$. The expected loss conditional on X, with the model is:

$$L_x^W = E_i\{\ell[|i - P_1(X)|] \mid X\} = P_1(X)\ell[1 - P_1(X)] + [1 - P_1(X)]\ell[P_1(X)] = \psi[P_1(X)]$$

The unconditional expected loss with the model is

$$L^W = E_X\{\psi[P_1(X)]\} \tag{23}$$

Note that $E_X[P_1(X)] = f_1$.

It makes sense to define the goodness-of-fit measure, g, to be the fractional reduction in expected loss from the introduction of the model:

$$g = 1 - (L^W/L).$$

But before one can do that, it is important to find which loss functions result in values of g that range from zero to one; for otherwise, the loss function would not be reasonable.

Theorem 2: If Theorem 1 holds, g ranges from zero to one.

Proof: First we show the $L^W \in [0, L]$. Note that (a) $L^W \geq 0$ because $\psi(x) \geq 0$, and (b) $L^W = E_X\{\psi(P_1(X))\} \leq \psi\{E_X[P_1(X)]\} = \psi(\bar{f}_1) = L$ because, by virtue of theorem 1, $\psi(\cdot)$ is concave. Thus, $L^W \in [0, L]$.

That the limit values 0 and L can be reached is clear because if $P_1(X)$ is constant, it must equal f_1, and $L_X^W = L^W = L$. On the other hand, if $P_1(X)$ takes values differing from zero and one by less than ε except for a set of X-values with probability measure less than ε, $L_X^W \to 0$ as $\varepsilon \to 0$ except for a set of X-values of measure zero. Thus, as $\varepsilon \to 0$, $L^W \to 0$, and g can be made arbitrarily close to one by letting $P_1(\cdot)$ resemble a simple function, $\delta(\cdot)$, sufficiently well.

It follows that any value of g in the [0, 1] interval can be attained with a convex combination of $P_1(X) = f_1$ and $P_1(X) = \delta(X)$. QED

The quadratic, logarithmic, and bang-bang losses yield three well-known goodness-of-fit measures.

(A) Quadratic Loss, g_q

From Equations 12 and 20 and by using the quadratic loss formula in the latter, we see that $L \equiv MSE$. Similarly, Equations 15 and 23 yield that $L^W \equiv MSE^W$. Thus, $g_q = d^2$, and the quadratic loss function generates the d-measure.

(B) Logarithmic loss, g_ℓ

Replacing the logarithmic loss function, $\ell(x)$, into Equations 20 and 23, we find that the expected loss, L, is

$$L = -f_1 \log f_1 - (1 - f_1)\log(1 - f_1),$$

and that L^W is

$$L^W = E_X\{-p_1(X)\log P_1(X) - [1 - P_1(X)]\log P_1(X)\}.$$

If, as stated earlier, the distribution of X used represents its sampling distribution (it is assumed that sampling is not choice-based), this expression is the negative expected value of log-likelihood function (for one observation):

$$\mathcal{L}(X) = \log p_1(X) \quad \text{if } i = 1$$
$$= \log[1 - P_1(X)] \quad \text{if } i = 0,$$

and

$$L^W = E_X[\mathcal{L}(X)].$$

Thus, L^W/L represents the reduction in log-likelihood that is achieved by introduction of the model and g_ℓ is the square of the pseudo-corre-

lation coefficient that is commonly used [1, 2].

(C) Bang-bang loss, q_b

Repeating the steps with the new $\ell(x)$, we find that

$$L = \min(f_1, 1 - f_1),$$

and

$$L^W = E_X\{\min[P_1(X), 1 - P_1(X)]\}.$$

The quantity, $\min\{P_1(X), 1 - P_1(X)\}$, represents the fraction of times that we will be wrong if for a choice-maker with attributes X we predict choice 1 if $P_1(X) > 0.5$ and choice 0 if $P_1(X) < 0.5$. Thus, L^W represents the "% wrong" measure that is used with success tables, and q_b is the reduction in the percentage of "wrong" predictions that is achieved by introduction of the model.

It should be clear that other goodness-of-fit measures can be derived by proper redefinition of $\ell(x)$ and that different measures may be called for depending on the application. The discussion at the outset of this paper and in the section on the d-measure argued in favor of the d-measure when the application aim is the prediction of market shares for large groups of individuals.

The next subsection discusses the relative magnitudes of q_q, q_ℓ, and q_b and gives a formula relating q_q and q_ℓ when they are small.

RELATIONSHIP AMONG GOODNESS-OF-FIT MEASURES

Recall that L is the height of the curve on Figure 3 (up to a factor of 2) when the abscissa is f_1, and that L^W is the average of the height when the abscissa varies with $P_1(X)$. This scaling does not affect the value q_b that is ultimately obtained because q_b depends on the ratio L^W/L.

If the variation in the abscissa is small and f_1 is substantially different from (1/2), the bang-bang function will yield $L \simeq L^W$ (and $q_b \simeq 0$) since the function is linear over the relevant range of $P_1(X)$. On the other hand [and also for small $\text{var}\ P_1(X)$], if $f_1 = (1/2)$, L is 1 and $L^W = 1 - 2E_X[|P_1(X) - (1/2)|]$. Then, $q_b = 2E_X[|P_1(X) - (1/2)|]$. This illustrates that for the bang-bang function and for models with low variance for $P_1(X)$ (as usually occurs), q_b will be smaller when f_1 is substantially different from (1/2).

Since the quadratic and logarithmic losses yield smooth $\psi(x)$'s, we approximate L^W taking expectations on a two-term Taylor series expansion:

$$L^W \cong L + \psi''(f_1)\text{var}_X[P_1(X)]; \text{var}_X[P_1(X)] \to 0$$

$$g \cong |\psi''(f_1)|/L\ \text{var}_X[P_1(X)]; \text{var}_X[P_1(X)] \to 0 \quad (24)$$

Replacing in Equation 24 $\psi''(f_1)$ and L by the appropriate values we find

$$g_q/g_\ell \cong 2[-(1 - f_1)\log(1 - f_1) - f_1 \log f_1] = 2L_\ell; \text{var}_X[P_1(X)] \to 0 \quad (25)$$

In this expression, L_ℓ is the logarithmic loss without model. It is clear from Figure 3 that $q_q > q_\ell$ if f_1 is between (approximately) 0.2 and 0.8, and that $q_q < q_\ell$ otherwise. If the probabilities are very unbalanced and $\text{var}_X[P_1(X)]$ is small: $0 \simeq q_b < q_q < q_\ell$, but if $f_1 \simeq 0.5$, $q_\ell < q_q < q_b$. This is because if $f_1 \simeq 0.5$, q_b is comparable with $(\text{var}_X[P_1(X)])^{(1/2)}$ but q_q and q_ℓ are comparable with $\text{var}_X[P_1(X)]$.

While Equation 25 allows us to approximate d from the output of most computer programs with

$$d \cong \sqrt{g_q} \cong \sqrt{2[(\hat{\mathcal{L}} - \bar{\mathcal{L}})/N]} = \ell \quad (26)$$

where $\bar{\mathcal{L}}$ is the background log-likelihood [3], $\hat{\mathcal{L}}$ is the maximum log-likelihood, and N is the sample size, it is recommended to calculate d exactly by simple modification of the computer programs.

Equation 26, thus, suggests that estimated log-likelihood values can actually be used to assess prediction errors. It should be remembered, however, that the approximation is only valid when $(\text{var}_X[P_i(X)])^{(1/2)}$ is so small that values of $P_i(X)$ outside the range where the two-term Taylor series expansion of $\psi(x)$ is a good approximation are rare. For this to happen, the d-measure must be small, which happens quite often in practice.

EXAMPLE

Assume that we have the following binary probit model:

$$P_1 = 1 - P_2 = \Phi(\theta X),$$

where Φ is the standard normal c.d.f., and θ and X are scalars. Furthermore, assume that the sampling distribution of X, F(x), is normal with zero mean and variance one.

As the value of θ increases, most of the individuals in the population face choices that can be predicted with less and less uncertainty because unless $X = 0$, p_1 and p_2 approach either zero or one as $\theta \to \infty$. In the limit choices can be predicted deterministically and the d-measure is 1.

This example illustrates this phenomenon. It will calculate $d(\theta)$ and show how it increases from zero to one as θ goes from zero to ∞. It will also demonstrate that for small values of θ, the approximation discussed in the preceding section holds.

Let us first calculate f_i and $\text{var}_X[P_i(\theta, X)]$. Letting $\phi(x)$ denote the standard normal p.d.f., we write

$$f_i = \int_{-\infty}^{\infty} \Phi(\theta x)\phi(x)dx = \int_{-\infty}^{\infty} \Phi(-\theta x)\phi(x)dx$$

$$= \int_{-\infty}^{\infty} [1 - \Phi(\theta x)]\,\phi(x)dx = 1 - \int_{-\infty}^{\infty} \Phi(\theta x)\phi(x)dx$$

$$= 1 - f_i \Rightarrow f_i = 0.5$$

Also,

$$\text{var}_X[P_i(\theta, X)] = \int_{-\infty}^{\infty} \Phi^2(\theta x)\phi(x)dx - f_i^2$$

Consequently,

$$d(\theta) = [4\int_{-\infty}^{\infty} \Phi^2(\theta x)\phi(x)dx - 1]^{(1/2)} \quad (27)$$

It is clear that if $\theta = 0$, $d = 0$ and that $\lim_{\theta \to \infty} d(\theta) = 1$. Figure 5 plots $d(\theta)$.

The approximation given by Equation 26 can also be calculated:

$$(1/N)\,\hat{\mathcal{L}} = 2 \times 0.5 \log 0.5 = -0.693$$

$$(1/N)\,\hat{\mathcal{L}} = 2\int_{-\infty}^{\infty} \Phi(\theta x)\,[\log \Phi(\theta x)]\,\phi(x)dx$$

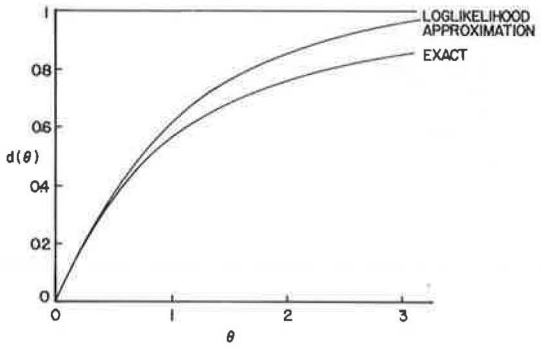
$$d(\theta) \approx [4\int_{-\infty}^{\infty} \Phi(\theta x)\,[\log \Phi(\theta x)]\,\phi(x)dx + 1.386]^{(1/2)}$$

Figure 5 also plots this equation, which, as expected, tracks well low-to-moderate correlation values.

CONCLUSION AND EXTENSIONS

The d-measure and the D-statistic have been shown to be the binary model equivalent of the correlation coefficient of regression in the sense that they can be related to predictive RMS error in the same way. It was shown that D was a consistent estimator for d and that for the typical low-correlation models en-

Figure 5. D-measure and its log-likelihood approximation.



countered in social sciences, d is in between the two other measures of goodness-of-fit that are commonly used: the likelihood ratio goodness-of-fit measure, g_ℓ, and the success table measure of goodness-of-fit, g_b.

If, as is usually the case, the goodness-of-fit measures are not high, it is possible to obtain an approximation for d that is based on the log-likelihood function, Equation 26. For models with low correlation, the measure ℓ itself may be regarded as a goodness-of-fit measure but care must be exercised because ℓ may exceed one.

Equation 26 is also useful because under the usual regularity conditions, $\hat{\ell}$ (the estimator for ℓ from the sample log-likelihood) is such that $(N\hat{\ell}^2)$ is χ^2 distributed with as many degrees of freedom as parameters. Of course, D has approximately the same distribution if small.

For models with more than two alternatives (i = 0,1,...,I-1), a prediction goodness-of-fit measure can be developed in the same way. In this instance, however, the goodness-of-fit measure is not the same for all the alternatives because $\text{var}_X[P_i(X)]$ is not the same for all. Consideration shows that the measure for the ith alternative, d_i, is

$$d_i = \sqrt{\{\text{var}_X[P_i(X)]\}/[f_i(1 - f_i)]},$$

and that a D-statistic can be derived in the same way. The interpretation in terms of prediction RMS error reduction (Figure 1) is unchanged.

ACKNOWLEDGMENT

The research reported in this paper was supported by NSF grants CEE-7812004 and CEE-81-11681 to the University of California, Berkeley.

REFERENCES

1. D. McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. In Frontiers of Econometrics (P. Zarembka, ed.), Academic Press, New York, 1970, pp. 105-142.
2. J.R. Hauser. Testing the Accuracy, Usefulness and Significance of Probabilistic Choice Models: An Information-Theoretic Approach. Operations Research, Vol. 26, 1978, pp. 406-421.
3. C.F. Daganzo. Multinomial Probit: The Theory and Its Application to Demand Forecasting. Academic Press, New York, 1979.

Appendix

CONSISTENCY OF THE D-STATISTIC

Let us assume that a random, attribute-based sampling process was used to gather a sample for estimation of a binary choice model. We denote by $X^{(n)}$ the attribute vector of the nth observation, by $i^{(n)}$, the choice of said observation and by N the number of observations. We assume that there is a set of parameters, θ, that are entered in the specification of the model:

$$0 < p_I = G(\theta., X) = 1 - p_o < 1 \tag{A1}$$

and that for a true (unknown), θ_o, $G(\theta_o, X)$ coincides with $P_1(X)$.

Let $F_X(x)$ denote the sampling distribution of X and write $\sigma^2(\theta)$ for the variance of $G(\theta, X)$; $\sigma^2(\theta_o)$ represents the variance of the choice probabilities in the numerator of Equation 16b. We assume that some regularity conditions to make the MLE of θ_o, $\hat{\theta}$, consistent hold, and in particular that

(A) $G(\theta, X)$ is a continuously differentiable function of θ_o in a δ_o neighborhood of it;
(B) The sample space of X is bounded, and
(C) $0 < G(\theta, X) < 1$.

These three properties will be used in the consistency proof of D.

Let $S_N^2[...]$ denote the sample variance function of the N elements in brackets. Then, Property (C) ensures that $\sigma^2(\theta_o)$ is finite and the consistency of the sample variance statistic enables us to write:

$$\left\{[S_N^2[G(\Theta_o, X^{(1)}), \ldots, G(\Theta_o, X^{(N)})] - \sigma^2(\Theta_o)]\right\} \xrightarrow{p} 0 \tag{A2}$$

where $\xrightarrow{p}$ denotes the limit in probability of the sequence in braces as $N \to \infty$. Note that the $X^{(n)}$ are i.i.d. drawings from $F_X(x)$.

Because $\hat{\theta}$ is consistent and because of Equation A2, we can write for arbitrarily small values of ϵ', ϵ'', δ', and δ'':

$$\Pr\left\{|S_n^2[G(\theta_o, X^{(1)}), \ldots, G(\theta_o, X^{(n)})] - \sigma^2(\theta_o)| \leq \epsilon'\right\} \geq 1 - \delta'; \forall n'' \geq N'(\epsilon', \delta') \tag{A3}$$

and

$$\Pr\left[\| \hat{\Theta}(X^{(1)} \ldots X^{(n)}; i^{(1)} \ldots i^{(n)}) - \theta_o \| \leq \epsilon''\right] \geq 1 - \delta''; \forall n'' \geq N''(\epsilon'', \delta'') \tag{A4}$$

In these two expressions, the random variables $(X^{(1)} \ldots X^{(n)})$ are the same, $N'(\epsilon', \delta')$ and $N''(\epsilon'', \delta'')$ represent finite positive numbers that depend on (ϵ', δ') and (ϵ'', δ''), $\| \cdot \|$ represents the Euclidean norm, and $\hat{\theta}(...)$ is the function that relates the data, $(X^{(1)} \ldots X^{(n)}; i^{(1)} \ldots i^{(n)})$, to the MLE of θ_o.

Before proceeding, we need the following preliminary result:

Lemma 1

Under the regularity conditions previously stated, we can write for sufficiently small but positive values of (ϵ'', δ''):

$$\Pr\Big(|S_n^2\{G[\hat{\Theta}(X^{(1)} \ldots X^{(n)}; i^{(1)} \ldots i^{(n)}), X^{(1)}], \ldots, G[\hat{\Theta}(X^{(1)} \ldots X^{(n)}; i^{(1)} \ldots i^{(n)}), X^{(n)}]\} - S_n^2[G(\theta_o, X^{(1)}), \ldots, G(\theta_o, X^{(n)})]| \leq k\epsilon''\Big) \geq 1 - \delta''; \forall n'' \geq N''(\epsilon'', \delta'') \tag{A5}$$

for a positive, finite constant, k, which is independent of n.

The lemma states in words that the difference between the estimated and true sample variances converges uniformly in probability to zero as the sample size increases. This lemma in conjunction with (A3) will be used to show that the estimated sample variance converges in probability to $\sigma^2(\theta_o)$ (lemma 2). First we prove this lemma.

Proof: Let us rewrite for convenience the estimated sample variance as $\hat{S}^2_{n_2}$ and the sample variance for a given value of θ as $S^2_n(\theta|X)$. Because of property (A), $S^2_n(\theta|X)$ is a continuously differentiable function of θ with derivatives:

$$[\partial S^2_n(\theta|X)]/[\partial\theta_j]\,|_{\theta=\theta_o} = \sum_{i=1}^{n} (2/n)\Big\{G(\theta_o, X^{(i)}) - \sum_{m=1}^{n} G(\theta_o, X^{(m)}) \div n\Big\}\,[\partial G(\theta, X^{(i)})/\partial\theta_j]\,|_{\theta=\theta_o}$$

Because all of the elements in this sum are continuous for all values of $X = (X^{(1)},\ldots,X^{(n)})$, and by property (B) the $X^{(i)}$ values that can possibly occur are bounded, the addends are bounded by a number, M_j/n. Thus, the partial derivatives of $S^2_n(\theta|X)$ are uniformly bounded and for any θ in a small neighborhood of θ_o we can write:

$$|S^2_n(\theta|X) - S^2_n(\theta_o|X)| \leq \|\theta - \theta_o\|\, k; \|\theta - \theta_o\| < \delta'_o < \delta_o$$

for some $\delta'_o > 0$ and finite $k > 0$. The combination of this fact and (A4) proves the lemma (as long as $\epsilon'' \leq \delta'_o$). We now show that $\{\hat{S}^2_n\} \xrightarrow{p} \sigma^2(\theta_o)$.

Lemma 2

$$[\hat{S}^2_n] \xrightarrow{p} \sigma^2(\theta_o)$$

Proof: We shall prove that for any positive (ϵ, δ):

$$\Pr[|\hat{S}^2_n - \sigma^2(\theta_o)| \leq \epsilon] \geq 1 - \delta; \forall n \geq N(\epsilon, \delta) \qquad (A6)$$

where $N(\epsilon, \delta)$ is a finite positive integer that depends on ϵ and δ. Combining (A3) and (A5) we can write:

$$\begin{aligned}\Pr[|\hat{S}^2_n - \sigma^2(\theta_o)| \leq \epsilon' + k\epsilon''] &\geq \Pr[|\hat{S}^2_n - S^2_n(\theta_o|X)| \leq k\epsilon'' \text{ and } |S^2_n(\theta_o|X) - \sigma^2(\theta_o)| \leq \epsilon'] \\ &\geq 1 - (\delta' + \delta''), \text{ if } n \geq \max[N'(\epsilon', \delta'); N''(\epsilon'', \delta'')] \qquad (A7)\end{aligned}$$

It is now clear that Equation A6 follows from Equation A7 with $\delta' = \delta'' = \delta/2$, $\epsilon' = k\epsilon'' = \epsilon/2$, and $N(\epsilon, \delta) = \max\{N'[(\epsilon/2), (\delta/2)]; N''[(\epsilon/2k), (\delta/2)]\}$.

Lemma 3

Identical arguments show that the sample mean $\bar{G}_n$:

$$\bar{G}_n = (1/n)\sum_{i=1}^{n} G[\tilde{\theta}(X^{(1)},\ldots,X^{(n)}, i^{(1)},\ldots,i^{(n)}), X^{(i)}]$$

converges in probability to $f_1 = E_X[P_1(X)]$

Theorem: The D-statistic is a consistent estimator for d.

Proof: From the definition,

$$d = f[f_1, \sigma^2(\theta_o)] = \sqrt{\sigma^2(\theta_o)/f_1(1 - f_1)} \qquad (A8)$$

where the function $f(.,.)$ is continuous if $f_1 \epsilon (0, 1)$. Regularity condition (C) implies that $f_1 = E_X[G(\theta_o, X)]$ is in the open unit interval and therefore confirms the continuity of $f(.,.)$.

This continuity implies that if f_1 and $\sigma^2(\theta_o)$ are replaced by consistent estimators, $\bar{G}_n$ and $\hat{S}^2_n$, as the arguments of $f(.,.)$, the result:

$$f(\bar{G}_n, \hat{S}^2_n) = \sqrt{[\hat{S}^2_n/\bar{G}_n(1 - \bar{G}_n)]} = D$$

is a consistent estimator.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Evaluation of Usefulness of Two Standard Goodness-of-Fit Indicators for Comparing Non-Nested Random Utility Models

JOEL L. HOROWITZ

The likelihood ratio index and the percentage of correctly predicted choices in the estimation sample are two well-known goodness-of-fit indicators for logit and other random utility models. They are used frequently for comparing non-nested models (i.e., models such that neither can be obtained from the other by choosing suitable values of the estimated parameters) to determine which best explains the available data. The results of an investigation of the abilities of the two statistics to distinguish between correct and incorrect models in such comparisons are reported. It is shown that with estimation data sets of practical size, a slightly modified form of the standard likelihood ratio index has good ability to distinguish between correct and incorrect models when the root-mean-square (RMS) difference between the two models' choice probabilities exceeds 10 to 15 percent. In addition, very small differences between the values of two models' modified likelihood ratio indices indicate with high probability that the model with the lower index value is incorrect. The percent-correctly-predicted statistic is considerably less useful for comparing models. It can fail to distinguish between correct and incorrect models whose choice probabilities differ by at least 25 percent (RMS), even with arbitrarily large estimation samples. Moreover, there are no readily available criteria for determining how large the differences between the values of two models' percent-correctly-predicted statistics must be to justify a conclusion that the model with the lower value likely is incorrect. Several travel-related examples are given.

Travel decisions frequently entail choices among discrete sets of alternatives, such as frequencies, destinations, modes, and routes of travel, and it

often is necessary in travel behavior research and practical transportation studies to be able to predict the outcomes of such choices. In recent years, multinomial logit and multinomial probit models have begun receiving extensive use for this purpose. These models are examples of a broader class of models, called random utility models, that are derived from the behavioral principle of utility maximization. In models of this class, it is assumed that an individual's preferences among the available alternatives can be described with a utility function and that the individual selects the alternative with the greatest utility. The utility of an alternative is represented as the sum of a deterministic and a random component. The deterministic component accounts for systematic effects of observed factors that influence choice whereas the random component accounts for the effects of unobserved factors. The random utility model then predicts the probability that a randomly selected individual with given values of the observed factors will choose a particular alternative (i.e., the probability that the utility of the particular alternative exceeds the utilities of all other alternatives). [See Domencich and McFadden (1) and Hensher and Johnson (2) for detailed discussions of the behavioral foundations of random utility models.]

A logit, probit, or other random utility model constitutes a functional relation between the observed factors (or explanatory variables) and the probabilities that an individual chooses the various alternatives. These relations usually are not known a priori and must be estimated by fitting a model to observations of choices and the explanatory variables. The fitting process usually takes place in two steps. [See McFadden (3) and Daganzo (4) for detailed discussions of the fitting process.] In the first step, the functional form of the relation between the explanatory variables and the choice probabilities is specified up to a finite set of constant parameters. For example, it might be specified that a mode choice model has the multinomial logit functional form with a utility function that is a linear combination of a certain set of travel time and cost variables. The coefficients of these variables in the linear utility function would then constitute the set of constant parameters. In the second step of the fitting process, the values of the parameters are estimated statistically from the observations. The method of maximum likelihood usually is used for this purpose. The statistical theory on which this two-step process is based assumes that the first step is carried out without error. If this assumption is true and certain mild regularity conditions are satisfied, then the maximum likelihood estimates of the parameter values have a variety of desirable statistical properties. Most importantly, the estimated parameter values and the values of the choice probabilities computed from the estimated parameters approach the true values as the size of the estimation data set increases toward infinity.

In practice, of course, the correct functional forms of the relations between the choice probabilities and the explanatory variables are not known a priori, even up to a set of constant parameters. Not surprisingly, use of an incorrect functional form in the second estimation step can lead to models that produce highly erroneous forecasts of travel behavior (5-7). Consequently, the development of empirical random utility models, like the development of most other statistically based models, usually includes comparing several models with different functional forms in an effort to distinguish forms that are likely incorrect from ones that may be correct.

This paper is concerned with comparisons that involve two models. Several procedures are available for carrying out such pairwise comparisons, depending on whether the models being compared are nested or non-nested. Two models are nested if one model can be obtained from the other by assigning appropriate values to the latter model's parameters. In non-nested models, this cannot be done; given the values of either model's parameters, it is not possible to choose values of the other model's parameters so that the two models become identical. (Examples of nested and non-nested models are given in the following section of this paper.) Comparisons of nested models usually are carried out by using likelihood ratio or t-tests (3). The statistical properties of these tests are well known, and numerical experiments with the tests have indicated that they have good ability to distinguish between correct and seriously erroneous random utility models in the nested case (6,7).

Comparisons of non-nested models usually are made with so-called goodness-of-fit statistics, such as the likelihood ratio index and the percentage of choices correctly predicted if each individual in the estimation data set is assumed to choose the alternative with the highest probability in the estimated model. The basis for using these statistics for comparing models is largely intuitive. Plausibly, it is assumed that a correct model is likely to have larger values of these statistics than are incorrect models. Thus, models with relatively large values of these statistics are presumed more likely to be correct (or to be less erroneous) than models with relatively low values of these statistics. However, there has been no systematic investigation of the abilities of these statistics to distinguish reliably between correct and incorrect models. This paper reports the results of such an investigation. It is shown that with estimation data sets of practical size, a slightly modified form of the standard likelihood ratio index statistic has good ability to distinguish between correct and incorrect models when the root-mean-square (RMS) difference between the choice probabilities of the two models exceeds 10-15 percent. The percent-correctly-predicted statistic is considerably less reliable. It can fail to distinguish between models whose choice probabilities differ by at least 25 percent (RMS), even if the size of the estimation data set is allowed to increase without bound.

EXAMPLES OF NESTED AND NON-NESTED MODELS

The results presented in this paper apply only to comparisons of non-nested models. Serious errors can result from attempting to apply the results in the nested case. Therefore, it is important to understand the distinction between nested and non-nested models. This distinction is illustrated by the following examples. [See Horowitz (8) for a precise mathematical definition of non-nestedness.]

Example 1 (Nested Models)

Suppose that two logit models are being considered for possible use in forecasting mode choice. For simplicity, assume that there are only two modes, automobile and transit. The two models differ in the specifications of their utility functions. (In the following discussion the term "utility function" will refer to the deterministic component unless otherwise noted.) The utility function for mode i (i = automobile or transit) in model 1 is

$$U_i = \alpha_1 (\text{Time}_i) + \beta_1 (\text{Cost}_i) \quad (1)$$

where

U_i = utility of mode i,
$Time_i$ = travel time of mode i,
$Cost_i$ = travel cost of mode i, and
α_1 and β_1 = constant parameters.

In model 2 the utility function is

$$U_i = \alpha_2 (Time_i) \quad (2)$$

In either model, the probability of choosing mode i is given by the binomial logit function

$$P_i = \exp(U_i)/[\exp(U_1) + \exp(U_2)] \quad (3)$$

where P_i is the choice probability. Then model 2 is nested with model 1, since model 2 can be obtained from model 1 by setting $\alpha_1 = \alpha_2$ and $\beta_1 = 0$.

Example 2 (Non-Nested Models)

Suppose that in the same mode choice study a third model (model 3) is being considered and that its utility function for mode i is

$$U_i = \alpha_3 \log(Time_i) + \beta_3 \log(Cost_i) \quad (4)$$

As in models 1 and 2, the choice probabilities in model 3 are related to the utility function by Equation 3. Then models 1 and 3 and models 2 and 3 form non-nested pairs. Apart from the degenerate case in which all of the parameter values are zero, it is not possible to choose parameter values for models 1 and 3 so that the two models coincide, nor is it possible to choose parameter values for models 2 and 3 so that those models coincide.

CRITERIA FOR EVALUATING COMPARISON PROCEDURES

In this paper the likelihood ratio index and percent-correctly-predicted goodness-of-fit statistics will be evaluated according to their abilities to distinguish between correctly and incorrectly specified models. Two factors must be taken into account in making these evaluations: the abilities of the statistics to distinguish between correct and incorrect models in the absence of random sampling error, and the effects of random sampling error on the comparisons. Random sampling error arises because different individuals with the same observable characteristics (i.e., the same values of a model's explanatory variables) and the same sets of alternatives may make different choices, owing to the effects of unobserved factors (i.e., the random component of the utility function). As a result, the estimated parameter values, choice probabilities, and goodness-of-fit statistics for a model tend to have different values in different finite samples of individuals, even if the model involved is correctly specified. These random fluctuations in estimation results can cause a goodness-of-fit statistic for an incorrectly specified model to be more favorable than that for a correctly specified model. Random sampling error, therefore, constitutes a "noise factor" that impairs the ability of test statistics to distinguish correct models from incorrect ones.

Random sampling error always can be made insignificantly small by making the estimation sample for a model sufficiently large. Moreover, if the estimation sample is large enough to make the effects of sampling error insignificant, then it always is possible to determine unambiguously if a model is correct by comparing the values of its choice probabilities for each set of values of the explanatory variables with the observed choices of individuals with the same values of the explanatory variables. A model whose choice probabilities for the available alternatives differ from the observed proportions of individuals choosing these alternatives is incorrect. Accordingly, it is reasonable to demand of comparison statistics, such as the goodness-of-fit statistics discussed here, that they be capable of distinguishing without error between correct and incorrect models in the absence of random sampling error. In formal statistical terms, this property of a test is called consistency. Statistical test procedures that are not consistent usually are considered to be unacceptable.

In practice, of course, it usually is not possible to work with samples sufficiently large to eliminate the effects of random sampling error. As has already been noted, random sampling error can cause comparison procedures to give misleading results (e.g., to cause an incorrectly specified model to have a more favorable goodness-of-fit statistic than a correctly specified one), even if the procedures always would give correct results in samples large enough to eliminate sampling error. In general, it is not possible to guarantee that a comparison procedure always will distinguish correctly between correct and incorrect models when sampling error is present, particularly if the choice probabilities of the two models would not be greatly different in the absence of sampling error. However, to be useful, a comparison procedure should be capable of making the distinction correctly most of the time (i.e., with high probability) if the two models would have significantly different choice probabilities in the absence of sampling error.

The foregoing discussion suggests that to be useful, a test procedure should have the following two characteristics:

1. It should be consistent. In other words, in the absence of random sampling error it should always distinguish correctly between correctly and incorrectly specified models.
2. It should not be highly sensitive to random sampling error. In other words, in samples of practical size (typically 100-1000 observations) the procedure should have a high probability of rejecting an incorrect model in a comparison with a correct one if the two models would yield substantially different values of the choice probabilities in the absence of sampling error.

It also is desirable that the value of the test statistic associated with a comparison procedure be easy to compute. However, the two procedures discussed in this paper have roughly equal computational requirements so that computational considerations do not provide a basis for a comparative evaluation of the procedures.

In the following two sections, the likelihood-ratio-index and percent-correctly-predicted statistics will be evaluated according to the foregoing two criteria.

LIKELIHOOD RATIO INDEX

The most commonly used form of the likelihood ratio index, and the only form that will be discussed here, is defined as follows. [See Daganzo (4) and Tardiff (9) for definitions and discussions of other forms.] Let L denote the value of a model's log-likelihood function when the values of the model's parameters equal their maximum likelihood estimates. Let L_0 denote the value of the log-likelihood function of a model that assigns equal values to the choice probabilities of all alternatives, regardless of the values of the explanatory variables. Then

the likelihood ratio index, ρ^2, is defined as

$$\rho^2 = 1 - L/L_0 \quad (5)$$

If there are N individuals in the estimation sample and each individual chooses among J alternatives, then L_0 is given by

$$L_0 = -N \log J \quad (6)$$

(Throughout this paper it will be assumed that all individuals face the same number of alternatives. Allowing different individuals to face different numbers of alternatives would add complexity to the presentation without changing the results significantly.)

The likelihood ratio index is a goodness-of-fit statistic for random utility models that is similar in many respects to the coefficient of multiple determination, R^2, in regression models. The larger the value of ρ^2 for a model, the better the model fits the given data. Therefore, two nonnested models P and Q can be compared by comparing the likelihood ratio indices ρ_P^2 and ρ_Q^2 for the two models. If $\rho_P^2 - \rho_Q^2 > 0$, this suggests that model P is superior to model Q, whereas $\rho_P^2 - \rho_Q^2 < 0$ suggests that model Q is superior.

Suppose that model P is correct and model Q is incorrect. Then, to evaluate the likelihood ratio index according to the criteria given in the previous section it is necessary to determine (a) if $\rho_P^2 - \rho_Q^2 > 0$, always, when the sample size N is large enough to make the effects of random sampling error insignificant and (b) if $\rho_P^2 - \rho_Q^2 > 0$ is a high probability outcome in samples of practical size if models P and Q would yield substantially different values of the choice probabilities in the absence of random sampling error. These determinations can be made if the probability distribution of $\rho_P^2 - \rho_Q^2$ is known.

The following notation will be used in describing the probability distribution of $\rho_P^2 - \rho_Q^2$ and in the subsequent discussion. Let P(i,X) denote the true probability that an individual chooses alternative i when the explanatory variables have the value X. (Here, X denotes the entire set of values of all of the explanatory variables of both models. If some elements of X are not variables of one of the models, then the choice probabilities of this model are independent of the values of these elements.) Let Q(i,X) denote the choice probability for alternative i that model Q would yield when the explanatory variables have the value X if there were no random sampling error. P(i,X) and Q(i,X), respectively, are the large-sample limits (i.e., the limits as the sample size approaches infinity) of the maximum likelihood estimates of the choice probabilities of models P and Q. Let $p_x(X)$ denote the proportion of individuals in the population being studied for whom the values of the explanatory variables equal X. Let k_P and k_Q, respectively, denote the numbers of estimated parameters in models P and Q. As before, let N denote the number of individuals in the estimation data set, and let J denote the number of alternatives available to each individual. Finally define Δ^2 by

$$\Delta^2 = \sum_{i,X} \left\{[P(i,X) - Q(i,X)]/P(i,X)\right\}^2 P(i,X)p_x(X) \quad (7)$$

Δ^2 is the weighted mean square fractional error in the choice probabilities that would result from using the incorrect probabilities Q(i,X) in place of the correct probabilities P(i,X). The weight for given i and X values equals the proportion of the population that has explanatory variable values X and selects alternative i.

Note that Δ^2 always exceeds zero unless Q(i,X) = P(i,X) for all i and X. In other words, Δ^2 exceeds zero unless model Q is identical to model P and, therefore, is correctly specified.

The probability distribution of $\rho_P^2 - \rho_Q^2$ is derived in Horowitz (8). It is shown there that $\rho_P^2 - \rho_Q^2$ has approximately the normal distribution with the following mean (μ) and variance (σ^2):

$$\mu = \Delta^2/2 \log J + (k_P - k_Q)/2N \log J \quad (8)$$

$$\sigma^2 = \Delta^2/N (\log J)^2 \quad (9)$$

The accuracy of the approximation increases as the sample size, N, increases. It follows from Equations 8 and 9 that $\rho_P^2 - \rho_Q^2$ exceeds zero, thereby indicating that the correct model is superior to the incorrect one, with the following probability:

$$\Pr(\rho_P^2 - \rho_Q^2 > 0) = \Phi\left[N^{1/2}\Delta/2 + (k_P - k_Q)/2N^{1/2}\Delta\right] \quad (10)$$

where Φ is the cumulative standard normal distribution function.

It is easy to see from Equation 10 that the likelihood ratio index satisfies the consistency criterion given in the previous section. As N approaches infinity, $\Pr(\rho_P^2 - \rho_Q^2 > 0)$ approaches 1. Since the limit of an infinite sample corresponds to eliminating random sampling error, this result implies that in the absence of sampling error, ρ_P^2 always exceeds ρ_Q^2. Thus, if there is no sampling error, the likelihood ratio index always indicates that the correct model is superior to the incorrect one.

It also can be seen from Equations 8-10 that with finite samples, adding parameters to an incorrect model (i.e., increasing k_Q) tends to decrease the value of $\rho_P^2 - \rho_Q^2$ and of $\Pr(\rho_P^2 - \rho_Q^2 > 0)$, even if the variables associated with the added parameters are incorrectly specified or irrelevant to the choices being studied. This clearly is an undesirable characteristic of the likelihood ratio index because it means that in finite samples the index tends to favor models with large numbers of parameters, regardless of whether these models are correct. However, this characteristic can be removed by making a simple modification in the definition of the likelihood ratio index. Define $\bar{\rho}^2$, the modified likelihood ratio index, for a model with k estimated parameters by

$$\bar{\rho}^2 = \rho^2 - k/2N \log J \quad (11)$$

or, equivalently,

$$\bar{\rho}^2 = 1 - (L - k/2)/N \log J \quad (12)$$

The modified statistic $\bar{\rho}^2$ is used in the same way as ρ^2 for comparing two models. Thus, $\bar{\rho}_P^2 - \bar{\rho}_Q^2 > 0$ indicates that model P is superior to model Q, and $\bar{\rho}_P^2 - \bar{\rho}_Q^2 < 0$ indicates that model Q is superior.

Equations 10 and 11 imply that the probability that $\bar{\rho}_P^2 - \bar{\rho}_Q^2$ exceeds zero is given by

$$\Pr(\bar{\rho}_P^2 - \bar{\rho}_Q^2 > 0) = \Phi(N^{1/2}\Delta/2) \quad (13)$$

It can be seen from Equation 13 that $\bar{\rho}^2$ is consistent and, in contrast to the unmodified likelihood ratio index, is not biased in favor of models with large numbers of parameters. This makes $\bar{\rho}^2$ more

Table 1. Probabilities that modified likelihood ratio index selects correct model (P) in comparison with incorrect model (Q).

N	Δ	$Pr(\bar{\rho}_P^2 - \bar{\rho}_Q^2 > 0)$	N	Δ	$Pr(\bar{\rho}_P^2 - \bar{\rho}_Q^2 > 0)$
100	0.05	0.60	250	0.15	0.88
	0.10	0.69		0.20	0.94
	0.15	0.77	500	0.05	0.71
	0.20	0.84		0.10	0.87
250	0.05	0.66		0.15	0.95
	0.10	0.79		0.20	0.99

Notes: N = size of the estimation data set.
Δ = RMS difference between the large sample limiting values of the choice probabilities of models P and Q.
$Pr(\bar{\rho}_P^2 - \bar{\rho}_Q^2 > 0)$ = probability that the correct model is selected.

useful than ρ^2 for comparing models. Accordingly, only the modified index $\bar{\rho}^2$ will be used in the remainder of this paper.

The performance of $\bar{\rho}^2$ according to the second criterion given in the previous section can be assessed by computing $Pr(\rho_P^2 - \rho_Q^2 > 0)$ for various values of N and Δ. Table 1 shows the results of such a computation. It can be seen that if the sample size exceeds roughly 250, a comparison of two models using $\bar{\rho}^2$ has a probability of at least 0.80 of selecting the correct model when the RMS percentage difference between the two models' choice probabilities (i.e., 100Δ) exceeds 10 to 15 percent.

The probability distribution of $\bar{\rho}_P^2 - \bar{\rho}_Q^2$ can be used to derive a simple upper bound on the probability that $\bar{\rho}^2$ for an incorrect model (Q) exceeds $\bar{\rho}^2$ for a correct model (P) by an arbitrary amount z. The bound is [see Horowitz (8)]

$$Pr(\bar{\rho}_Q^2 - \bar{\rho}_P^2 > z) \leq \Phi[-(2Nz \log J)^{1/2}] \quad (14)$$

This inequality implies that in moderate size samples, very small differences between the $\bar{\rho}^2$ values of two models indicate with high probability that the model with the lower $\bar{\rho}^2$ value is incorrect. For example, if $N \geq 250$, $z \geq 0.01$, and $J \geq 2$, inequality (15) yields

$$Pr(\bar{\rho}_Q^2 - \bar{\rho}_P^2 > z) \leq 0.03 \quad (15)$$

In other words, if $N \geq 250$ and the $\bar{\rho}^2$ values of two models differ by 0.01 or more, the model with the lower $\bar{\rho}^2$ value almost certainly is incorrect.

PERCENT-CORRECTLY-PREDICTED STATISTIC

The percent-correctly-predicted statistic for a model is obtained by "predicting" that each individual in the model's estimation data set chooses the alternative that has the highest choice probability according to the estimated model. The predictions are compared with the observed choices of the individuals in the estimation data set, and the percentage of correct predictions (i.e., the percentage of individuals for which the predicted and observed choices coincide) is computed. The result yields the percent-correctly-predicted statistic. In this section an example is presented in which the percent-correctly-predicted statistic fails to satisfy either of the previously defined evaluation criteria. The implications of the example for the usefulness of the percent-correctly-predicted statistic are discussed following the presentation of the example.

Notation and Formulas

The example is based on binomial logit models (i.e., logit models of choice between two alternatives). Before presenting the example, it is necessary to present the notation and formulas that will be used in computing the percent-correctly-predicted statistic for the example.

Let model P be correct and model Q be incorrect. As before, let P(i,X) denote the true probability that an individual with explanatory variables X chooses alternative i (i = 1 or 2), and let Q(i,X) denote the large-sample limit of the model Q estimate of the probability that this individual chooses alternative i. Let the number of individuals in the estimation sample be N, and let these individuals be indexed by n (n = 1, 2,..., N). Let $\hat{P}(i,X)$ and $\hat{Q}(i,X)$, respectively, denote the maximum likelihood estimates of the model P and model Q choice probabilities obtained from the estimation sample. Thus, for example $\hat{P}(i,X_n)$ and $\hat{Q}(i,X_n)$ are the estimated probabilities that individual n chooses alternative i by using models P and Q. Note that as N approaches infinity, $\hat{P}$ approaches P and $\hat{Q}$ approaches Q.

For each individual n in the estimation sample define j_n by

j_n = 1 if individual n was observed to choose alternative 1, 0 otherwise (16)

Thus, j_n indicates the observed choice of individual n. Also, for each individual n in the estimation sample define $\hat{S}_P(n)$ and $\hat{S}_Q(n)$ by

$$\hat{S}_P(n) = \begin{cases} 100 \text{ if } \hat{P}(1,X_n) > 1/2 \text{ and } j_n = 1 \text{ or } \hat{P}(1,X_n) < 1/2 \\ \quad \text{and } j_n = 0; \\ 50 \text{ if } \hat{P}(1,X_n) = 1/2; \text{ and} \\ 0 \text{ if } \hat{P}(1,X_n) > 1/2 \text{ and } j_n = 0 \text{ or } \hat{P}(1,X_n) < 1/2 \\ \quad \text{and } j_n = 1 \end{cases} \quad (17)$$

$$\hat{S}_Q(n) = \begin{cases} 100 \text{ if } \hat{Q}(1,X_n) > 1/2 \text{ and } j_n = 1 \text{ or } \hat{Q}(1,X_n) < 1/2 \\ \quad \text{and } j_n = 0; \\ 50 \text{ if } \hat{Q}(1,X_n) = 1/2; \text{ and} \\ 0 \text{ if } \hat{Q}(1,X_n) > 1/2 \text{ and } j_n = 0 \text{ or } \hat{Q}(1,X_n) < 1/2 \\ \quad \text{and } j_n = 1 \end{cases} \quad (18)$$

Then $\hat{S}_P(n)$ equals 100 if individual n was observed to choose the alternative with the larger value of $\hat{P}(i,X_n)$ (i = 1 or 2), $\hat{S}_P(n)$ equals 0 if individual n was observed to choose the alternative with the lower P value, and $\hat{S}_P(n) = 50$ if the two P values for individual n are equal. Thus, $\hat{S}_P(n)$ is the percent-correctly-predicted statistic for model P by using the single individual n. An analogous interpretation applies to $\hat{S}_Q(n)$. The percent-correctly-predicted statistics for models P and Q by using the entire estimation sample, $\hat{S}_P$ and $\hat{S}_Q$, are obtained by averaging $\hat{S}_P(n)$ and $\hat{S}_Q(n)$ over all of the individuals in the sample. Thus,

$$\hat{S}_P = (1/N) \sum_n \hat{S}_P(n) \quad (19)$$

$$\hat{S}_Q = (1/N) \sum_n \hat{S}_Q(n) \quad (20)$$

$\hat{S}_P$ and $\hat{S}_Q$ coincide with the usual definitions of the percent-correctly-predicted goodness-of-fit statistics for models P and Q.

The large-sample limits of $\hat{S}_P$ and $\hat{S}_Q$ can be computed by applying the strong law of large numbers to Equations 19 and 20. This yields the following result:

Table 2. Values of the explanatory variables and choice probabilities for the example.

ΔT (min)	ΔC (cents)	P (auto, ΔT)[a]	P (transit, ΔT)[a]	Q (auto, ΔC)[a]	Q (transit, ΔC)[a]
5	20	0.62	0.38	0.71	0.29
10	5	0.73	0.27	0.56	0.44
20	40	0.88	0.12	0.86	0.14

[a]P and Q are computed as large-sample limits and, therefore, are free of random sampling error. The large-sample limit of the maximum likelihood estimate of β is 0.045.

$$S_P = 100 \sum_x p_x(X) \max[P(1, X), P(2, X)] \quad (21)$$

$$S_Q = 100 \sum_x p_x(X) \left\{ I(X) P(1, X) + [1 - I(X)] P(2, X) \right\} \quad (22)$$

where S_P and S_Q, respectively, denote the large-sample limits of $\hat{S}_P$ and $\hat{S}_Q$ and I(X) is defined by

$$I(X) = \begin{cases} 1 \text{ if } Q(1, X) > 1/2; \\ 1/2 \text{ if } Q(1, X) = 1/2; \text{ and} \\ 0 \text{ if } Q(1, X) < 1/2 \end{cases} \quad (23)$$

Equation 21 has been derived previously by Daganzo (10). S_P and S_Q are the values that the percent-correctly-predicted statistics for models P and Q would have if there were no random sampling error.

Example

Suppose that two binomial logit models, P and Q, of mode choice between automobile and transit are being considered. In model P the probability of choosing automobile is given by

$$P(\text{auto}, \Delta T) = 1/[1 + \exp(-\alpha \Delta T)] \quad (24)$$

where Δ T is transit travel time minus automobile travel time, and α is a constant. In model Q the probability of choosing automobile is given by

$$Q(\text{auto}, \Delta C) = 1/[1 + \exp(\beta \Delta C)] \quad (25)$$

where Δ C is transit travel cost minus automobile travel cost and β is a constant. The transit choice probabilities are equal to one minus the automobile choice probabilities.

Suppose that model P is correct and that the true value of α is 0.10. Suppose also that in the population being studied there are only three possible combinations of values of Δ T and Δ C, as shown in Table 2, and that these combinations occur with equal probability. Thus $p_X(X) = 1/3$ for all values of the explanatory variables of both models. Given the values of α and Δ T, the values of the true choice probabilities P(i, Δ T) (i = automobile or transit) can be computed from Equation 25, and the large-sample-limit of $\hat{Q}$(i, Δ C) can be computed by using methods described in Horowitz (6). The results are shown in Table 2. It can be seen from the table that the differences between the true choice probabilities P and large-sample-limit model Q choice probabilities vary from 2 to 63 percent. The RMS percentage difference, as computed from Equation 7, is 25 percent. Thus, the true and erroneous models yield substantially different values of the choice probabilities in the absence of random sampling error.

The large-sample limits of the percent-correctly-predicted statistics for models P and Q can be computed from Equations 21 and 22. The result is $S_P = S_Q = 74$ percent. Thus, the large-sample limits of the percent-correctly-predicted statistics for the two models are equal. This means that when there is no random sampling error, these statistics provide no information useful for distinguishing between the correct model P and the incorrect model Q, even though there are large differences between the choice probabilities of the two models. Clearly, the percent-correctly-predicted statistic fails to satisfy the first of the previously defined evaluation criteria in this case.

It also can be shown that the percent-correctly-predicted statistic is not useful for distinguishing between models P and Q with finite samples. (This is to be expected because reducing the sample size from infinity reduces the information content of the sample.) With a finite sample, the percent-correctly-predicted statistics of models P and Q can differ only if there are values of Δ T and Δ C for which the alternative with the highest estimated model P choice probability is different from the alternative with the highest estimated model Q choice probability. In terms of the notation developed in the previous subsection, this means that there must be pairs of values of Δ T and Δ C such that either $\hat{P}$(auto, Δ T) > 0.5 and $\hat{Q}$(auto, Δ C) < 0.5 or $\hat{P}$(auto, Δ T) < 0.5 and $\hat{Q}$(auto, Δ T) > 0.5, as can be seen from Equations 18 and 19. The probability that either or both of these events occurs can be computed from the information in Table 2 by using methods described by Daganzo (4,11) and Horowitz (12). The result is that with an estimation data set of 100 or more observations, the probability that either or both of these events occurs is virtually zero (i.e., less than 5×10^{-6}). Thus, the percent-correctly-predicted statistics of models P and Q are virtually certain to be equal with any reasonable estimation sample size. It follows that the percent-correctly-predicted statistic almost certainly will fail to distinguish between the two models with any reasonably sized estimation sample. Thus, the statistic fails to satisfy the second of the previously defined evaluation criteria.

The performance of the percent-correctlypredicted statistic in this example may be contrasted with that of the modified likelihood ratio index. The Δ value for comparing models P and Q is 0.25. It follows from Equation 13 that $\bar{\rho}_P^2 - \bar{\rho}_Q^2$ exceeds zero with probability 0.89 if the sample size is 100 and probability 0.98 if the sample size is 250. The probability approaches 1.0 as the sample size increases further. Thus, in contrast to the percent-correctly-predicted statistic, the modified likelihood ratio index has a high probability of distinguishing correctly between models P and Q with any reasonably sized estimation sample.

Discussion of Example

The foregoing example shows that the percent-correctly-predicted statistic may fail to distinguish between a correct and an incorrect model, even if the choice probabilities of the two models differ substantially. Of course, this does not mean that the statistic always will fail to make such distinctions. On the contrary, it is possible to construct examples in which the statistic distinguishes between correct and incorrect models quite satisfactorily. However, the example shows that the statistic is an unreliable diagnostic tool. The fact that two models have similar values of the percent-correctly-predicted statistic does not necessarily

imply that the two models have similar abilities to explain the available data or are equally likely to be correct.

Although small differences between the values of two models' percent-correctly-predicted statistics do not necessarily imply that the two models are equally satisfactory, the possibility remains that a large difference between the values of the statistics implies with high probability that the model with the lower value is incorrect. This possibility clearly is fulfilled qualitatively. However, it does not appear to be possible to develop for the percent-correctly-predicted statistic an inequality analogous to inequality 14 for the modified likelihood ratio index. Thus, it does not appear possible, at least analytically, to specify quantitatively a minimum difference between the values of two models' percent-correctly-predicted statistics that enables one to conclude with high probability that the model with the lower value is incorrect. As was discussed in connection with inequality 14, it is possible to specify such a minimum difference for the modified likelihood ratio index. Although the apparent lack of a "minimum significant difference" for the percent-correctly-predicted statistic is disappointing and clearly impairs the statistic's usefulness, it should not be considered surprising or unusual. For example, there also is no known minimum significant difference for the R^2 values of two linear regression models.

CONCLUSIONS

The results presented here indicate that the modified likelihood ratio index provides a powerful method for comparing non-nested random utility models. It has been shown that very small differences between the values of the modified likelihood ratio indices of two models indicate with high probability that the model with the lower index value is incorrect. Moreover, if one of the models being compared is correct and there are substantial differences between the choice probabilities of the correct and incorrect models, the modified likelihood ratio index has a high probability of indicating that the correct model is superior to the incorrect one with estimation samples of practical size.

The percent-correctly-predicted statistic is much less useful for comparing models. The statistic may fail to distinguish between correct and incorrect models, regardless of sample size, even if the choice probabilities of the two models are very different. Moreover, there are no readily available quantitative criteria for determining how large the differences between the values of two models percent-correctly-predicted statistics must be to justify a conclusion that the model with the lower value likely is incorrect.

In summary, the modified likelihood ratio index is a much more useful tool for comparing non-nested random utility models than is the percent-correctly-predicted statistic.

ACKNOWLEDGMENT

The views expressed in this paper are mine. They are not necessarily endorsed by the U.S. Environmental Protection Agency.

REFERENCES

1. T.A. Domencich and D. McFadden. Urban Travel Demand: A Behavioral Analysis. American Elsevier Publishing Company, New York, 1975.
2. D.A. Hensher and L.W. Johnson. Applied Discrete-Choice Modelling. Croom-Helm, London, 1981.
3. D. McFadden. Analysis of Qualitative Choice Behavior. In Frontiers in Econometrics (P. Zarembka, ed.), Academic Press, New York, 1974.
4. C. Daganzo. Multinomial Probit: The Theory and Its Application to Demand Forecasting. Academic Press, New York, 1979.
5. J.L. Horowitz. Sources of Error and Uncertainty in Behavioral Travel Demand Models. In New Horizons in Travel-Behavior Research (P.R. Stopher, A.H. Meyburg, and W. Brog, eds.), Lexington Books, Lexington, MA, 1981.
6. J. Horowitz. Identification and Diagnosis of Specification Errors in the Multinomial Logit Model. Transportation Research, Vol. 15B, 1981, pp. 345-360.
7. J. Horowitz. The Accuracy of the Multinomial Logit Model as an Approximation to the Multinomial Probit Model of Travel Demand. Transportation Research, Vol. 14B, 1980, pp. 331-341.
8. J.L. Horowitz. Statistical Comparison of Non-Nested, Discrete-Choice, Random-Utility Models. U.S. Environmental Protection Agency, 1981.
9. T.J. Tardiff. A Note on Goodness-of-Fit Statistics for Probit and Logit Models. Transportation, Vol. 5, 1976, pp. 377-388.
10. C.F. Daganzo. Goodness-of-Fit Measures and the Predictive Power of Discrete Choice Models. TRB, Transportation Research Record 874, 1982, pp.13-19.
11. C.F. Daganzo. The Statistical Interpretation of Predictions with Disaggregate Demand Models. Transportation Science, Vol. 13, 1979, pp. 1-12.
12. J. Horowitz. Confidence Intervals for the Choice Probabilities of the Multinomial Logit Model. TRB, Transportation Research Record 728, 1979, pp. 23-30.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.