U.S. Department
of Transportation

**National Highway
Traffic Safety
Administration**

DOT-HS-807-446
DOT-TSC-NHTSA-89-3
Final Report

August 1989

Statistical Estimation of Rollover Risk

Peter Mengert
Santo Salvatore
Robert DiSario
Robert Walter

Transportation Systems Center
Office of Research and Analysis
Cambridge, MA 02142

Prepared for

Research and Development
Office of Crash Avoidance Research
Washington, DC 20590

This document is available to the public from the National Technical Information Service, Springfield, Virginia 22161

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

NOTICE

The United States Government does not endorse products of manufacturers. Trade of manufacturers'names appear herein solely because they are considered essential to the object of this report.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT-HS-807-446		

4. Title and Subtitle	5. Report Date
STATISTICAL ESTIMATION OF ROLLOVER RISK	August 1989
	6. Performing Organization Code: DTS-45
7. Author(s): P. Mengert, S. Salvatore, R. DiSario, R. Walter	8. Performing Organization Report No.: DOT-TSC-NHTSA-89-3
9. Performing Organization Name and Address: U.S. Department of Transportation, Research and Special Programs Administration, Transportation Systems Center, Cambridge, MA 02142	10. Work Unit No. (TRAIS): HS902/S9001
	11. Contract or Grant No.
12. Sponsoring Agency Name and Address: U.S. Department of Transportation, National Highway Traffic Safety Administration, Research and Development, Washington, DC 20590	13. Type of Report and Period Covered: Final Report, March 1988 - May 1989
	14. Sponsoring Agency Code: NRD-51

15. Supplementary Notes

16. Abstract

This report describes the results of a statistical analysis to determine the probability of a rollover in a single vehicle accident. Over 39,000 accidents, which included 4910 rollovers in the states of Texas, Maryland, and Washington were examined for 40 vehicle make/models using logistic regression analyses. Mathematical models were developed that related vehicle factors such as wheelbase and stability factor, (one-half the track width divided by the center of gravity height) and accident factors, (driver and the environmental variables), to rollover probability.

It was found that at the accident level (predicting rollover versus nonrollover) the vehicle stability factor and whether the accident occurred on an urban or rural road are important predictors of rollover probability. At the vehicle make/model level (comparison of predicted and actual rollover rates for the 40 make/models), the index of agreement (r^2) exceeded 0.90 with the stability factor in the regression model. Without the stability factor, the r^2 dropped to 0.53 indicating the importance of this factor. Other factors added little to predicting rollover at the make/model level.

17. Key Words	18. Distribution Statement
Rollover, Logistic Regression, Single Vehicle Accidents	DOCUMENT IS AVAILABLE TO THE PUBLIC THROUGH THE NATIONAL TECHNICAL INFORMATION SERVICE, SPRINGFIELD, VIRGINIA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	72	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

PREFACE

This report describes a statistical estimate of the risk of a vehicle rolling over when involved in a single vehicle accident. With increasing sales of utility vehicles, vans, and trucks, more vehicles with higher centers of gravity are on the road. It is important for NHTSA to better understand the relationship of the vehicle design factors and the driver and the environment in a rollover accident. This report is part of that effort to understand this complex mix of factors that can contribute to such an accident.

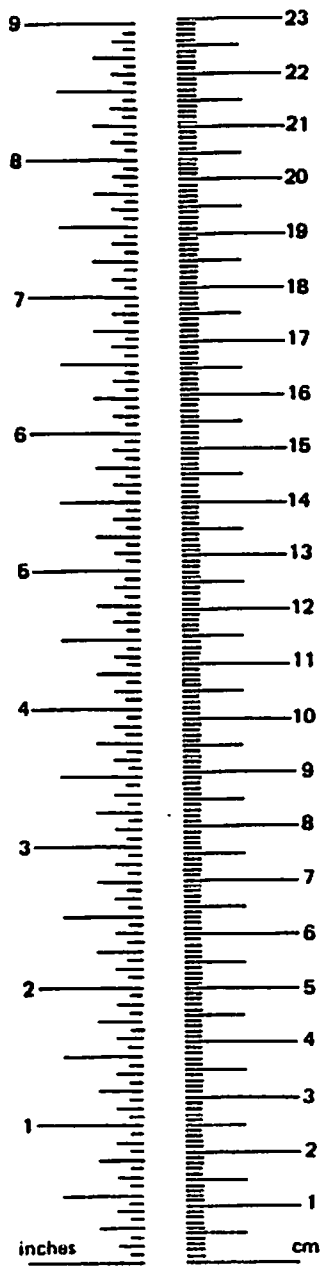
This effort was sponsored by the U.S. Department of Transportation, National Highway Traffic Safety Administration, Office of Research and Development. The support and technical advice of Dr. H. Keith Brewer, Chief of the Light Vehicle Dynamics and Simulation Division and Anna Harwin of his division are gratefully acknowledged.

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 In. = 2.54 cm (exactly). For other exact conversions and more detail tables see NBS Misc. Publ. 286, Units of Weight and Measures. Price $2.25 SD Catalog No. C13 10 286.



Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	36	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F

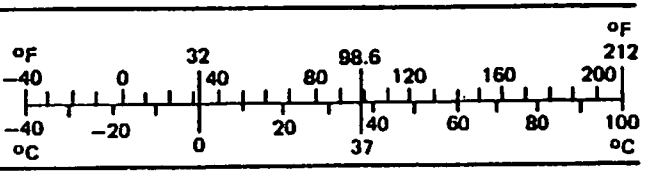


TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The purpose of this paper is to determine the probability of a rollover (RO) in a single vehicle accident (SVA) as a function of both accident and vehicle variables. The method used was the analysis of 39,956 accidents in the states of Texas, Maryland, and Washington (of these, 4910 accidents were rollovers). The analyses used 40 make/models of both passenger cars and utility vehicles.* Using logistic regression techniques, mathematical models were developed relating vehicle factors, including wheelbase (WB), and stability factor (SF) (one-half track width divided by center of gravity height), and accident factors (driver and environment) to rollover probability. These models contained between 7 and 11 predictor variables (see Table ES-1).

The results of the analyses were examined at the accident level (predicting rollover versus nonrollover) and at the make/model level (comparison of predicted and actual rollover rates for each of the 40 make/models). Both levels of analysis are summarized below.

Accident Level

- The vehicle stability factor and land-use variable (urban/rural) are important predictors of whether or not the accident resulted in rollover. Other variables such as the driver's age and sex contribute less to predicting rollover.

- The predictive power of the stability factor changed little with the inclusion or exclusion of all non-vehicle, variables including the urban/rural variable, indicating that the effect of SF is not due to confounding with any of these variables.

Make/Model Level

- r^2 (an index of the agreement between the actual and (model predicted rollover rates) exceeds 0.90 with the stability factor in the regression model (see Figure ES-1).

- The r^2 drops to 0.53 without the stability factor (see Figure ES-2).

*Note: The term "make/model" refers to vehicle type, while "model" alone will refer to a statistical or mathematical model.

TABLE ES-1. VARIABLES INCLUDED IN 11-FACTOR MODEL

Variable	Description
Stability Factor (SF)	1/2 Track width/center of gravity height
Wheelbase (WB)	Distance between front and rear tires
Rural	Accident occurred in rural vs. urban setting
Durban	Rural variable present vs. missing
Curve	Accident occurred on straight vs. curved road
Driver Error	Error vs. no error
Stable	Tracking vs. skidding, spinning
Youth	Driver age less than 25 vs. 25 and older
Alcohol and Drug Use	Driver under the influence, vs. not
Belt	Seat belt used vs. not used
Surf	Road surface dry vs. wet, snow, ice

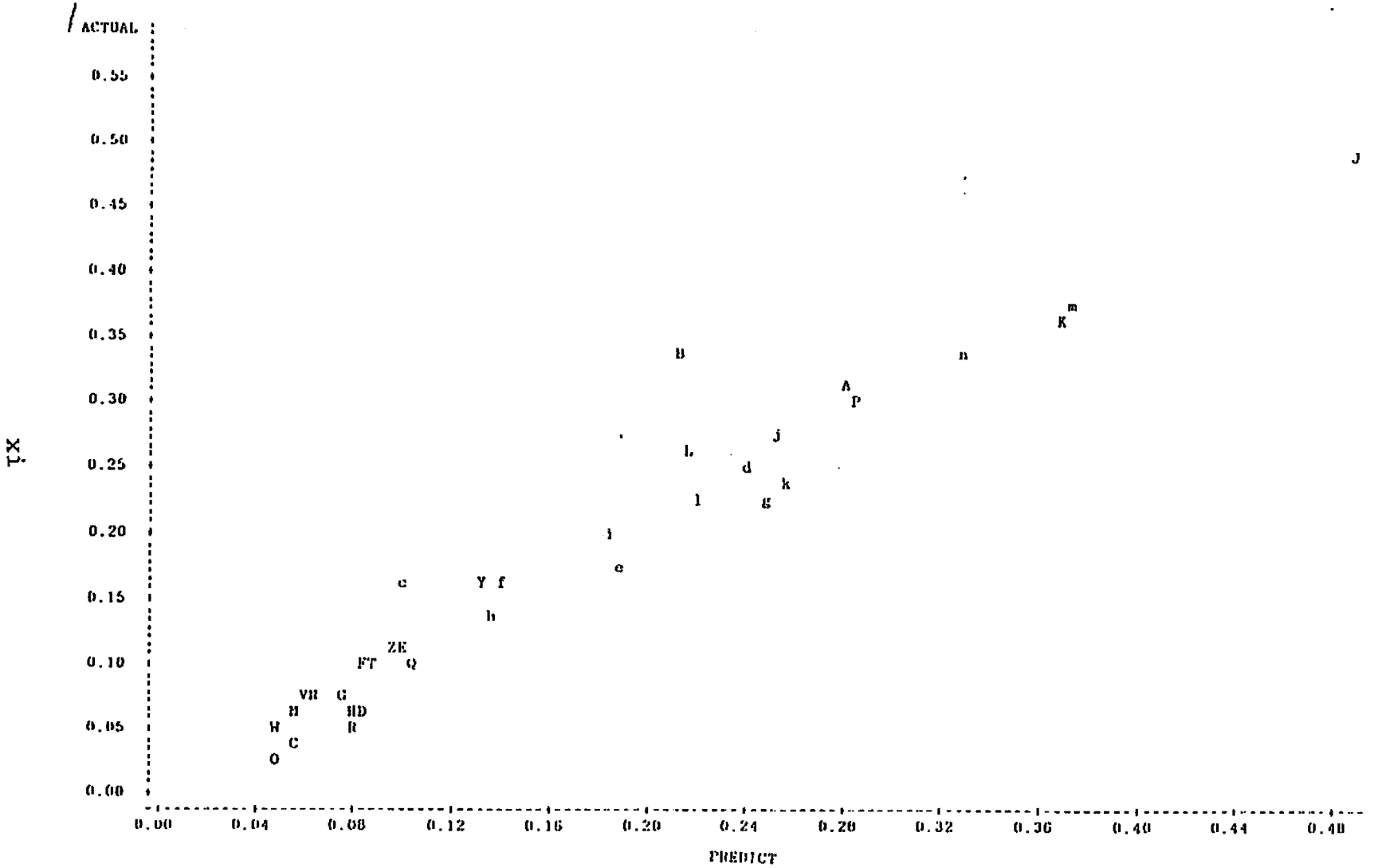


FIGURE ES-1. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR BASIC 11-FACTOR MODEL INCLUDING SF

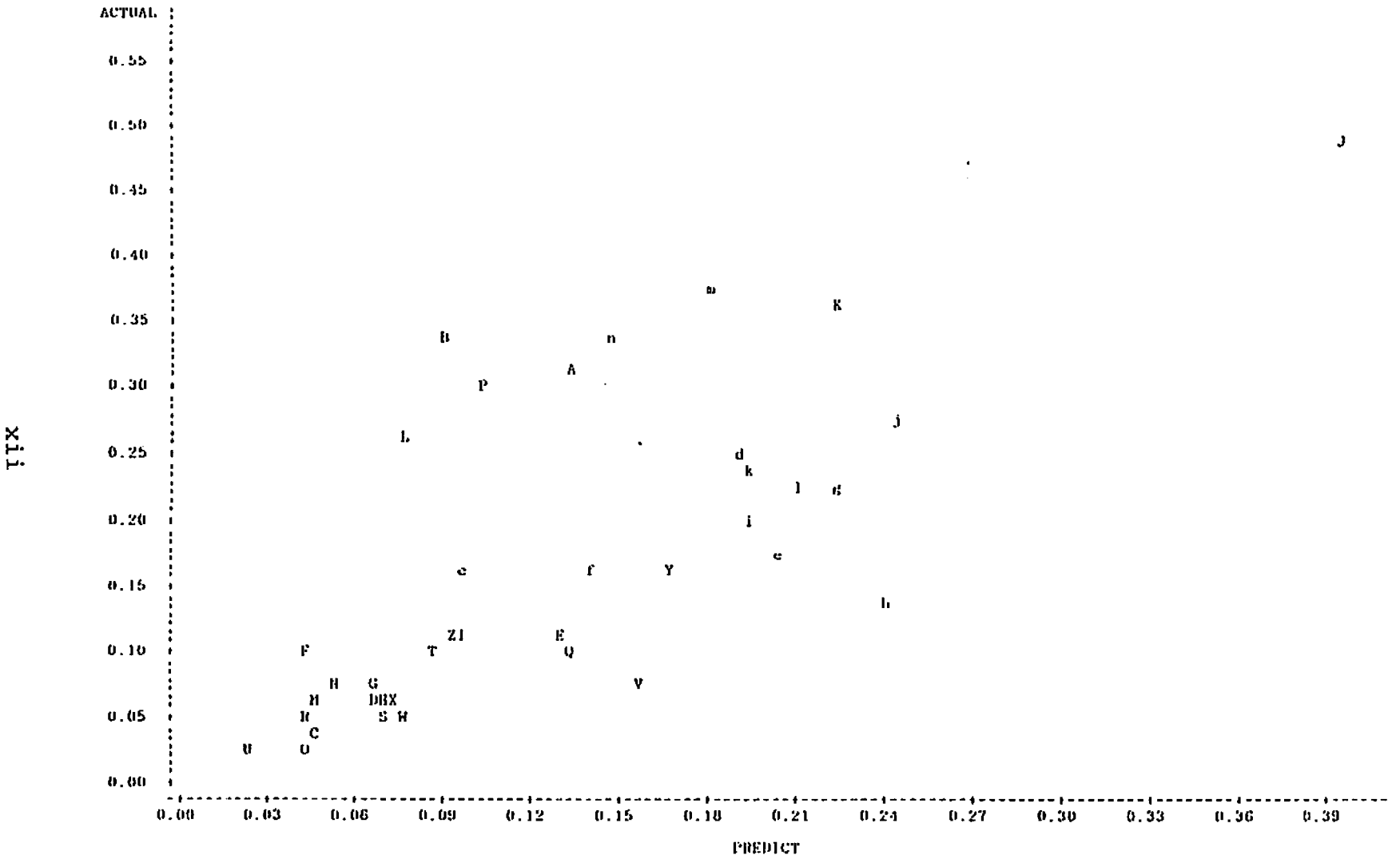


FIGURE ES-2. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR BASIC 11-FACTOR MODEL EXCLUDING SF

- The importance of the stability factor is enhanced at the make/model level, whereas the importance of the land-use variable is reduced.

- Nonvehicle factors are of little use for predicting rollover at the make/model level.

From these results, it was concluded that a mathematical model can be constructed that is an accurate predictor of the probability that a vehicle will roll over during a single vehicle accident. At both the accident and make/model level, the derived model was highly influenced by vehicle geometric factors, especially the stability factor. Other variables such as whether the accident occurred in the city or country or if a driver error was involved were also of importance, especially at the accident level. However, the effect of these environmental or driver variables was diminished at the make/model level.

TECHNICAL SUMMARY

PROBLEM

Studies have indicated that vehicles with a high center of gravity and narrow track width are involved in a disproportionate number of rollover (RO) single vehicle accidents (SVAs).[1,2] Utility vehicles intended for off-road operation characteristically incorporate such geometric design features. The popularity of the utility vehicle, which can be driven off-road as well as on-road, has prompted both Congressional and National Highway Traffic Safety Administration (NHTSA) action to further examine this issue. Studies of the risk of vehicle rollover use accident data to compare utility vehicles with passenger vehicles. These studies attempt to construct mathematical models using the accident data that predict a vehicle's rollover potential during a single vehicle accident based upon vehicle properties and accident variables.* These models are usually developed with linear regression techniques. The most recent studies by Robertson and Kelly[1] and by Harwin and Brewer,[2] using linear regression techniques, developed models that indicated that vehicle factors are the most important indicators of rollover potential in a single vehicle accident. Specifically, the stability factor (SF), which is defined as the ratio of one-half the tread width to the center of gravity height, was most highly correlated with rollover rates in single vehicle accidents. The addition of other "accident factors" relating to the accident environment or the driver did not significantly improve the ability of the model to predict rollovers. NHTSA's review of these results suggested that improvements could be made in these analyses by the use of logistic regression techniques. This paper reports on the results of that analysis.

APPROACH

The U.S. Department of Transportation, Research and Special Programs Administration, Transportation Systems Center (TSC) performed logistic regression analyses using the accident data previously analyzed by Harwin and Brewer. The accident data was derived from the Crash Avoidance Research Database (CARDfile) which contains the police accident reports from six states -

*This is usually an estimate of the fraction of single vehicle accidents which result in rollover. All the analyses in this report are based on single vehicle accidents and rollovers as a subset of these accidents.

Texas (TX), Maryland (MD), Washington (WA), Indiana (IN), Michigan (MI), and Pennsylvania (PA). From over two million accidents in TX, MD, and WA (1983 to 1985), 39,956 single vehicle accidents were analyzed. This analysis used 40 make/models of utility vehicles, and domestic and imported passenger cars with known stability and other vehicle factors (see Table TS-1*). In the data-set, the stability factor varied from 1.01 to 1.57.** The data contained 4910 rollovers. The ratio of ROs to SVAs varied by make/model from 0.021 to 0.489. Using both the Statistical Analysis System (SAS) and the Biomedical Data Package (BMDP) at the National Institutes of Health computer facility, we performed logistic regression analyses with single vehicle accidents which involved the selected vehicles and their related accident variables. Mathematical models were developed that contained both vehicle factors (stability factor and wheelbase) and accident variables relating to the driver, the vehicle, and the environment. The complete list of CARDfile variables and those used in the analysis can be found in Table TS-2. During a preliminary analysis, those variables that were most highly correlated with rollover were identified. These variables were then used in various combinations to predict the actual rollover experience during a single vehicle accident.

RESULTS

The main results of the logistic regression analyses are given in Table TS-3. These results show that at the accident level the variables which are very useful in predicting the probability of rollover in single vehicle accidents are SF, wheelbase, land-use (rural/urban), and driver error. However, the primary importance of the stability factor is seen as the result of several observations:

1. Leaving SF and WB out of a large 11-factor model lowered the likelihood ratio (as measured by the LIS explained in Section 3) more than leaving out all variables except SF (compare LIS of 781 for the former case with 1109 for the latter case).

2. Leaving SF out of the 11-factor model lowered the LIS much more than leaving out WB (LIS= 1478 vs. LIS=1856).

*Note: Tables TS-1, TS-2, and TS-3 are identical to Tables 3, 4, and 5, respectively, in the body of this report.
**These values were obtained from References 2 and 7.

TABLE TS-1. VEHICLE DATA

Vehicle No.	Make/ Model	Year	Wheel B. (inches)	Cg. Ht. (inches)	Tread W. /2 in.	S.F.	Data Contents Roll Ov. Acc.	Data Contents Sgl.Veh. Acc.	Ratio RO/SVA
1	JEEP CJ-5	1972-75	83.5	26.5	26.8	1.01	88	180	.489
2	JEEP CJ-7	1983-85	93.5	26.3	27.9	1.06	89	248	.359
3	JEEP CHEROKEE	1975-83	108.7	26.6	30.7	1.15	17	64	.266
4	FORD BRONCO	1973-83	104.7	27.9	30.0	1.08	216	710	.304
5	CHEVY BLAZER S-10	1983	100.5	27.2	29.6	1.09	77	246	.313
6	CHEVY BLAZER	1982	106.5	28.1	32.6	1.16	30	89	.337
7	TOYOTA LANDCRUISER	ALL	98.0	27.7	29.1	1.05	75	197	.381
8	INTER. H. SCOUT	<1979	100.0	27.7	29.3	1.06	86	257	.335
9	CAD. DEVILLE/BROUGHAM	1981 - 84	121.4	21.7	30.6	1.41	5	161	.031
10	CHEVY CITATION	1980 - 84	104.9	21.0	28.9	1.38	132	1197	.110
11	OLDS. OMEGA	1980 - 81	104.9	21.0	28.9	1.38	21	182	.115
12	BUICK SKYLARK	1980 - 84	104.9	21.0	28.9	1.38	31	328	.095
13	PONTIAC PHOENIX	1980 - 84	104.9	21.0	28.9	1.38	43	267	.161
14	CHEVY CHEVETTE	1979	97.3	19.8	27.0	1.36	46	273	.168
15	CHEVY CORVETTE	1973	98.0	18.2	28.6	1.57	3	40	.075
16	CHEVY CAMARO	ALL	108.0	18.7	29.4	1.57	353	7156	.049
17	PONTIAC FIREBIRD	ALL	108.0	18.7	29.4	1.57	192	3585	.054
18	CHEVY MALIBU	1978 - 81	108.0	21.7	30.4	1.40	82	1171	.070
19	OLDS. CUTLASS	1978 - 81	108.0	21.7	30.4	1.40	130	2272	.057
20	CHEVY MONTE CARLO	1978 - 81	108.0	21.7	30.4	1.40	108	1659	.065
21	BUICK CENTURY/REGAL	1978 - 81	108.0	21.7	30.4	1.40	74	1401	.053
22	PONTIAC LEMANS	1978 - 81	108.0	21.7	30.4	1.40	20	355	.056
23	CHRYSLER CORDOBA	1977 - 81	114.7	20.3	29.9	1.47	21	509	.041
24	DODGE MIRANDA	1977 - 81	114.7	20.3	29.9	1.47	1	48	.021
25	DODGE DIPLOMAT	1977 - 81	112.7	20.8	29.9	1.44	13	187	.070
26	CHRYSLER LEBARON	1977 - 81	112.7	20.8	29.9	1.44	22	377	.058
27	FORD MUSTANG	1979 - 81	100.4	20.0	28.3	1.42	209	1869	.112
28	MERCURY CAPRI	1979 - 81	100.4	20.0	28.3	1.42	58	587	.099
29	FORD LTD	1979 - 81	114.0	21.2	28.4	1.34	148	1461	.101
30	MERCURY MARQUIS	1979 - 81	114.0	21.2	28.4	1.34	11	204	.054
31	AMC CONCORD	1980	108.0	19.6	26.7	1.36	36	549	.066
32	AUDI 4000	ALL	99.8	20.4	26.9	1.32	19	120	.158
33	DATSUN Z, ZX	ALL	91.3	19.4	27.2	1.40	283	2060	.137
34	DATSUN B210	ALL	92.1	20.3	24.2	1.19	551	2433	.226
35	RENAULT LE CAR	ALL	95	20.9	24.5	1.17	27	113	.239
36	HONDA CIVIC	<1983	94.5	20.7	26.3	1.27	335	1654	.203
37	TOYOTA COROLLA	<1979	93.3	20.7	25.5	1.23	307	1388	.221
38	VW BEETLE	<1980	94.5	22.5	26.6	1.18	588	2404	.245
39	VW RABBIT	ALL	94.5	21.1	27	1.28	288	1686	.171
40	MAZDA GLC	<1980	91.1	20.5	24.6	1.20	75	269	.279
					Totals		4910	39956	.123

TABLE TS-2. VARIABLE DEFINITIONS

MODEL VARIABLE	CARDFILE FILENAME	CARDFILE VARIABLE	VARIABLE VALUES	FREQUENCY	CATEGORY	CARDFILE SUBCATEGORIES INCLUDED DICHOTOMIZED MODEL VARIABLES
ALCAD	DRIVER	ALC-DRUG	-1	31886	NO USE	NO INDICATION, MISS, UNK.
			1	8070	USE	ALCOHOL, DRUGS
BELT	DRIVER	RESTRAIN	-1	32838	NO BELT	NOT USED, NOT EQUIP, MISS, UNK
			1	7118	BELT	USED
CLIMATE	ACCIDENT	WEATHER	-1	32147	CLEAR/CLOUD	CLEAR, CLOUDY, MISS, UNK
			1	7809	OTHER	RAIN, SNOW/ICE, OTHER
CURVE	ACCIDENT	ROAD-ALIG	-1	29224	STRAIGHT	STRAIGHT, MISS, UNK
			1	10732	CURVED	CURVED
DURBAN	ACCIDENT	LAND-USE	-1	22280	MISSING	MISSING, UNK
			1	17676	URB/RUR	URBAN, RURAL
HERR	DRIVER	DRIVER-ERR	-1	10463	NOERROR	NONE, MISS, UNK,
			1	29493	ERROR	SPEED,SIGN/SIGNAL,PASSING,ASLEEP,ETC.
PROFILE	ACCIDENT	ROAD-PRO	-1	32968	LEVEL	LEVEL, MISS, UNK
			1	6988	GRADE	GRADE
ROADLOC	ACCIDENT	IMPILOC	-1	8624	ON ROAD	ON ROADWAY, MISS, UNK
			1	31332	OFF ROAD	ON SHOULDER, OFF ROADWAY
RURAL	ACCIDENT	LAND-USE	-1	32412	URBAN	URBAN, MISS, UNK
			1	7544	RURAL	RURAL
SEXY	DRIVER	SEX	-1	13065	FEMALE	FEMALE
			1	26891	MALE	MALE, MISS, UNK
STABLE	VEHICLE	PRESTAB	-1	34557	STABLE	TRACKING, NOT APPLICABLE, MISSING, UNK
			1	5399	NONSTABLE	SKIDDING, SPINNING, JACKKNIFING
STEER	VEHICLE	AVOID	-1	37142	NO AVOID	NO AVOIDANCE, MISS, UNK,
			1	2814	AVOID	AVOID VEHICLE, PEDESTRIAN, ETC
SURF	ACCIDENT	ROAD-SUR	-1	27848	DRY	DRY, MISS, UNK
			1	12108	ICY	WET, SNOW/ICE, OTHER
YOUTH	DRIVER	AGE	-1	18206	OLD	25 AND OVER
			1	21750	YOUNG	LESS THAN 25

	RANGE	MIN/MAX	MEAN	SD
SF	.56	1.01	1.38	.147
		1.57		
WB	37.9	83.5	102.81	7.3
		121.4		

TABLE TS-3. LOGISTIC REGRESSION MODELS FOR ROLLOVER PROBABILITY CONDITIONAL ON SINGLE VEHICLE ACCIDENT

MODEL FACTOR	SF ONLY	7 F	7F REDUCED	11 F	11F-SF	11F-SF-WB	11F-WB
1 SF	-4.903800	-3.965900	-4.483400	-4.090000	*****	*****	-4.934200
	(45.41)	(28.70)	(21.20)	(29.37)	*****	*****	(44.52)
2 WB	*****	-.029990	-.030302	-.028107	-.080502	*****	*****
	*****	(10.83)	(8.437)	(10.06)	(36.47)	*****	*****
3 RURAL	*****	.456340	.449650	.456570	.475520	.481750	.452140
	*****	(19.00)	(13.22)	(18.94)	(20.03)	(20.68)	(18.77)
4 OURBAN	*****	-.393490	-.443740	-.397440	-.392100	-.293670	-.369840
	*****	(17.13)	(13.44)	(17.03)	(17.02)	(13.08)	(15.99)
5 CURVE	*****	.26478	.25394	.26075	.242860	.254560	.26693
	*****	(15.34)	(10.45)	(15.00)	(14.20)	(15.20)	(15.37)
6 HERR	*****	.404720	.407270	.392950	.375640	.374080	.397090
	*****	(18.01)	(13.51)	(17.26)	(16.65)	(16.79)	(17.45)
7 STABLE	*****	.215000	.265380	.287710	.286430	.297370	.289080
	*****	(9.59)	(8.675)	(12.07)	(12.25)	(13.01)	(12.14)
8 YOUTH	*****	*****	*****	.066667	.018644	.026348	.085073
	*****	*****	*****	(4.02)	(1.146)	(1.652)	(5.15)
9 ALCAD	*****	*****	*****	.093045	.079714	.070307	.094859
	*****	*****	*****	(4.71)	(4.11)	(3.71)	(4.812)
10 BELT	*****	*****	*****	.094840	.084718	.109700	.101900
	*****	*****	*****	(4.63)	(4.20)	(5.56)	(4.983)
11 SURF	*****	*****	*****	-.170520	-.148740	-.142570	-.173010
	*****	*****	*****	(8.82)	(7.85)	(7.68)	(8.95)
CONSTANT	4.620500	6.598100	7.318400	6.664400	6.488000	-1.6256	4.9588
	(32.56)	(29.36)	(22.37)	(29.47)	(29.21)	(49.9)	(33.41)
r-squared (Make/ Model Based)	.907	.9448	.9467	.9444	.5272	.6812	.9296
LIS (See Text)	1109	1832	*****	1907	1478	781	1856

3. Although SF and WB are collinear (r = 0.64) and tend to proxy for each other in predicting rollover probability, there is evidence in the coefficients that the major predictive power is in SF. This is because the coefficient of SF shrinks by only 17% upon the introduction of WB (from -4.934 to -4.090) while the coefficient of WB shrinks by almost a factor of 3 on the introduction of SF (from - 0.0805 to -0.0281).

4. The coefficient of SF does not shrink on the introduction of all nonvehicle variables. It changes only by a trivial amount: from -4.904 to -4.934.

A regression analysis that excluded Texas accidents (56% of the cases) indicated that the importance of the land-use variable (rural/urban) was underestimated in the previous models as Texas had no land-use variable (see Appendix C). This result indicated that land-use may be as important as SF in predicting rollover at the accident level. However, of more importance in the evaluation of SF as a predictor of rollover is the fact that the regression coefficient of SF changed little when land-use was added to or taken out of the model. Moreover, the predictive capability of land-use is greatly reduced at the make/model level and will not affect the conclusions given below.

With regard to the performance of the models with predicted and actual rollover rates aggregated to the make/model level, the primary importance of stability factor is accentuated.

1. The vehicle make/model r^2 of the model with SF only is far higher than that for the model with all other factors (compare 0.907 to 0.5272).

2. Several plots discussed in the body of the text show that any model containing stability factor predicts rollover rate at least fairly well and any model which does not contain SF predicts rollover rate very poorly.

3. The larger models containing both WB and SF lead to exceptionally accurate predicted rollover rates.

When the distributions of predicted probabilities based on actual and nominal data are observed, there is confirmation of the importance of SF in predicting rollover rate. There is also evidence that with regard to the influence on predicted rollover rates, the nonvehicle variables are remarkably well balanced over make/models.

1.0 BACKGROUND

Recent studies by Robertson and Kelly[1] and Harwin and Brewer[2] using statistical regression analysis, indicated that a vehicle's propensity to rollover is directly related to a "stability factor" (SF). The stability factor is defined as the ratio of one-half the track width to the center of gravity height.

Based partly on the Robertson-Kelly study, Congressman T. Wirth (D-CO) petitioned the National Highway Traffic Safety Administration (NHTSA) to establish a rule, based on the stability factor, to limit a vehicle's rollover potential (Congressman Wirth proposed a stability factor of 1.2 as being the minimally acceptable level). He also requested that NHTSA further study this issue, open a defect investigation, and warn the public of this potential problem. The major parts of this petition were denied, based in part on the limitations of the Robertson-Kelly study as well as the need for more evidence of the connection between rollover and stability factor and the need to study the role of other vehicle parameters in this question. The Robertson-Kelly limitations included the use of 14 make/models which tended to cluster the data and the use of the Fatal Accident Reporting System (FARS) data which made the results applicable to fatal accidents only. Harwin and Brewer improved on the Robertson and Kelly study by using 40 make/models and approximately 40,000 single vehicle accidents (SVAs), including but not limited to fatals, from the Crash Avoidance Research Database (CARDfile). Their results indicated a strong relationship between the stability factor and rollover accidents. An internal NHTSA review agreed that this study was a significant improvement over the previous study, but suggested that the number of observations was insufficient for the number of predictors that were tested. It was also suggested that a logistic regression be performed where each single vehicle accident would be treated as an observation rather than the vehicle make/model as the observation. The dependent variable would be rollover. Logistic regression lends itself well to analysis when using a dichotomous dependent variable such as rollover/nonrollover.

NHTSA requested that TSC enhance the Harwin-Brewer study by performing the logistic regression. This report details the results of an analysis of the relationship of the stability factor to rollover propensity using logistic regression analysis at the individual accident level.

Note: The term "make/model" refers to the vehicle type, while "model" alone will refer to a statistical or mathematical model.

2.0 TECHNICAL APPROACH

The approach that TSC used was to restructure the Harwin-Brewer CARDfile data on single vehicle accidents so that a logit analysis could be performed at the accident level. The Harwin-Brewer database that contained all SVAs, including rollovers, from the states of Maryland and Texas for 1984 and 1985 and Washington for 1983, 1984, and 1985 was used. Other predictors were also used in addition to the stability factor. These included those available from CARDfile relating to the driver, the vehicle, and the accident together with other variables relating to the vehicle geometry.

2.1 CARDfile ACCIDENT DATABASE

The Crash Avoidance Research Database (CARDfile) was developed by NHTSA to define problem areas and support research in crash avoidance. The police accident reports from the states of Texas, Maryland, Washington, Pennsylvania, Indiana, and Michigan are assembled into a common format in a Statistical Analysis System (SAS) structure. CARDfile had approximately four million accidents from these states for 1983 through 1985 that were available for analysis (see Table 1). (CARDfile for 1986 is now available and will be used in future analyses.) The CARDfile database is subdivided into three subfiles relating to the accident, the driver, and the vehicle. The data elements in each of these files is shown in Table 2. Another study has indicated that CARDfile is representative of both national demographics and the accident experience.[3] For a more detailed description of the CARDfile database, the reader is referred to the studies by Harwin and Brewer[2] and Edwards.[4]

The SVAs for Texas and Maryland for 1984 and 1985 and Washington for 1983 through 1985 were extracted from CARDfile for 40 make/models. These make/models were selected based on the availability and range of their stability factors and to represent a selection of passenger cars and utility vehicles, both domestic and imported. The selected makes and models, their geometry and stability factors, and the counts of SVAs of each vehicle from the selected states for each year are shown in Table 3, along with the number of rollovers for each vehicle. The 40 make/models were composed of 20 passenger cars and eight utility vehicles. Also included were 12 vehicles built on the identical body-line that also share many common body parts such as the Chevrolet Citation and the Pontiac Phoenix. The model years ranged from 1972 through 1985, and the stability factor from 1.01 to 1.57. The final data-set contained 39,956 SVAs of which 4910 were rollovers (ratio of rollovers to SVAs = 0.1229). On a

TABLE 1. SUMMARY OF CARDfile STATE CRASH EXPERIENCE (1983-1985)

STATE	NO. CRASHES	NO. VEHICLES
Indiana	480,399	854,571
Maryland	384,450	717,284
Michigan	1,023,366	1,724,288
Pennsylvania	414,210	694,854
Texas	1,341,415	2,326,103
Washington	338,307	617,093
Totals	3,982,147	6,934,193

TABLE 2. CARDfile DATA ELEMENTS

Accident File

Case ID (CASE)
State of Crash (STATE)
Day of Crash (DD)
Month of Crash (MM)
Year of Crash (YY)
*Accident Type (ACC-TYPE)
Time of Crash (TIME)
*Roadway Alignment (RD-ALIGN)
*Number of Vehicles Involved (NO-VEH)
*Location of Primary Impact (IMPILOC)
Intersection Characteristics (INT-CHAR)
*Weather Conditions (WEATHER)
*Road Surface (RD-SUR)
*Land-Use (LAND-USE)
Primary Impact (IMPACT 1)
Crash Severity (ACC-SEV)
Light Conditions (LIGHT)
Relation to Intersection (INT-REL)
*Roadway Profile (RD-PRO)
Roadway Separation (RD-SEP)

Vehicle File

Case ID (CASE)
Vehicle Number (VEH)
Vehicle Impact Number (VATYPE)
Vehicle ID Number (VIN)
Component Failure (FAILCOMP)
Fatally Injured Occupants (FATAL)
Possible Injury Occupants (POS-INJ)
Unknown Occupant Injury Severity (UNK-OCC)
Incapacitating Injury Occupants (INCAP)
Nonincapacitating Injury Occupants (NONINCAP)
State of Crash (STATE)
*Make/Model Code (MAKE-MOD)
*Model Year (MOD-YR)
Vehicle Type (VEH-TYPE)
*Pre Crash Stability (PRE-STAB)
*Avoidance Attempt (AVOID)
Uninjured Occupants (UN-INJ)

Driver File

Case ID (CASE)
State of Crash (STATE)
Vehicle Number (VEH)
*Driver Age (AGE)
*Alcohol/Drug Use (ALC-Drug)
*Restraint Use (RESTRAIN)
Helmet Use (HEL-OP)
*Driver Sex (SEX)
*Driver Error (DR-ERROR)

*Indicates use in logistic regression

TABLE 3. VEHICLE DATA

Vehicle No.	Make/ Model	Year	Wheel B. (inches)	Cg. Ht. (inches)	Tread W. /2 in.	S.F.	Data Contents Roll Ov. Acc.	Data Contents Sgl.Veh. Acc.	Ratio RO/SVA
1	JEEP CJ-5	1972-75	83.5	26.5	26.8	1.01	88	180	.489
2	JEEP CJ-7	1983-85	93.5	26.3	27.9	1.06	89	248	.359
3	JEEP CHEROKEE	1975-83	108.7	26.6	30.7	1.15	17	64	.266
4	FORD BRONCO	1973-83	104.7	27.9	30.0	1.08	216	710	.304
5	CHEVY BLAZER S-10	1983	100.5	27.2	29.6	1.09	77	246	.313
6	CHEVY BLAZER	1982	106.5	28.1	32.6	1.16	30	89	.337
7	TOYOTA LANDCRUISER	ALL	98.0	27.7	29.1	1.05	75	197	.381
8	INTER. H. SCOUT	<1979	100.0	27.7	29.3	1.06	86	257	.335
9	CAD. DEVILLE/BROUGHAM	1981 - 84	121.4	21.7	30.6	1.41	5	161	.031
10	CHEVY CITATION	1980 - 84	104.9	21.0	28.9	1.38	132	1197	.110
11	OLDS. OMEGA	1980 - 81	104.9	21.0	28.9	1.38	21	182	.115
12	BUICK SKYLARK	1980 - 84	104.9	21.0	28.9	1.38	31	328	.095
13	PONTIAC PHOENIX	1980 - 84	104.9	21.0	28.9	1.38	43	267	.161
14	CHEVY CHEVETTE	1979	97.3	19.8	27.0	1.36	46	273	.168
15	CHEVY CORVETTE	1973	98.0	18.2	28.6	1.57	3	40	.075
16	CHEVY CAMARO	ALL	108.0	18.7	29.4	1.57	353	7156	.049
17	PONTIAC FIREBIRD	ALL	108.0	18.7	29.4	1.57	192	3585	.054
18	CHEVY MALIBU	1978 - 81	108.0	21.7	30.4	1.40	82	1171	.070
19	OLDS. CUTLASS	1978 - 81	108.0	21.7	30.4	1.40	130	2272	.057
20	CHEVY MONTE CARLO	1978 - 81	108.0	21.7	30.4	1.40	108	1659	.065
21	BUICK CENTURY/REGAL	1978 - 81	108.0	21.7	30.4	1.40	74	1401	.053
22	PONTIAC LEMANS	1978 - 81	108.0	21.7	30.4	1.40	20	355	.056
23	CHRYSLER CORDOBA	1977 - 81	114.7	20.3	29.9	1.47	21	509	.041
24	DODGE MIRANDA	1977 - 81	114.7	20.3	29.9	1.47	1	48	.021
25	DODGE DIPLOMAT	1977 - 81	112.7	20.8	29.9	1.44	13	187	.070
26	CHRYSLER LEBARON	1977 - 81	112.7	20.8	29.9	1.44	22	377	.058
27	FORD MUSTANG	1979 - 81	100.4	20.0	28.3	1.42	209	1869	.112
28	MERCURY CAPRI	1979 - 81	100.4	20.0	28.3	1.42	58	587	.099
29	FORD LTD	1979 - 81	114.0	21.2	28.4	1.34	148	1461	.101
30	MERCURY MARQUIS	1979 - 81	114.0	21.2	28.4	1.34	11	204	.054
31	AMC CONCORD	1980	108.0	19.6	26.7	1.36	36	549	.066
32	AUDI 4000	ALL	99.8	20.4	26.9	1.32	19	120	.158
33	DATSUN Z, ZX	ALL	91.3	19.4	27.2	1.40	283	2060	.137
34	DATSUN B210	ALL	92.1	20.3	24.2	1.19	551	2433	.226
35	RENAULT LE CAR	ALL	95	20.9	24.5	1.17	27	113	.239
36	HONDA CIVIC	<1983	94.5	20.7	26.3	1.27	335	1654	.203
37	TOYOTA COROLLA	<1979	93.3	20.7	25.5	1.23	307	1388	.221
38	VW BEETLE	<1980	94.5	22.5	26.6	1.18	588	2404	.245
39	VW RABBIT	ALL	94.5	21.1	27	1.28	288	1686	.171
40	MAZDA GLC	<1980	91.1	20.5	24.6	1.20	75	269	.279
					Totals		4910	39956	.123

make/model basis, this ratio varied from 0.021 for the Dodge Mirada to 0.489 for the Jeep CJ-5.

2.2 DATA STRUCTURING

The data-set for this analysis was the final data-set used by Harwin and Brewer. It contained the 39,956 single vehicle accidents referred to in the previous section. CARDfile data tapes at the NIH computer facility in Bethesda, MD, were accessed remotely from TSC. Computer programs were made available by Harwin-Brewer and modified to suit program goals. They were used to extract the required accidents from CARDfile. The accident frequency, across states and years, for the 40 make/models selected for the study, were obtained and compared with the figures provided in the Harwin-Brewer report. Complete agreement was obtained. The 39,956 cases aggregated in this initial data-set were then transferred, in total, to the Managed Storage System at NIH. The advantages of the Managed Storage System are: (1) much more rapid access to the data-set than that provided by tape, and (2) considerably less storage cost than computer system hard disk.

The independent variables used in the data analysis (derived from CARDfile variables) are shown in Table 4. Also shown are the frequencies for the dichotomized variables and descriptive measures for the two continuous variables. To date, up to 16 independent variables have been entered as possible predictors of the dependent variable, rollover, in a single vehicle accident. As the table indicates, 14 of the independent variables are entered as dichotomies and the two vehicle geometry variables, stability factor and wheelbase, are entered as continuous variables.

For each dichotomous variable, Table 4 displays the frequencies of the two levels of the variable and, in the column adjacent to the frequency, a short descriptive title to assist the reader in comprehending the category. Two columns give the SAS variable name used by CARDfile and the file in which it may be found in the CARDfile database. The final column provides a quick list of the CARDfile variable values that were collapsed into model variables shown in the first column. The range, mean, and standard deviation of the two continuous variables are shown at the bottom of the table.

Note that the variable "DURBAN" like "URBAN" is based on the CARDfile "LAND-USE." This variable was needed because the majority of records had missing values for LAND-USE. It turns out that "DURBAN" which is by definition synonymous with missing LAND-USE, equals -1 for TEXAS and +1 for other states. These variables are discussed more completely in Appendix C.

TABLE 4. VARIABLE DEFINITIONS

MODEL VARIABLE	CARDFILE FILENAME	CARDFILE VARIABLE	VARIABLE VALUES	FREQUENCY	CATEGORY	CARDFILE SUBCATEGORIES INCLUDED DICHOTOMIZED MODEL VARIABLES
ALCAD	DRIVER	ALC-DRUG	-1	31886	NO USE	NO INDICATION, MISS, UNK.
			1	8070	USE	ALCOHOL, DRUGS
BELT	DRIVER	RESTRAIN	-1	32838	NO BELT	NOT USED, NOT EQUIP, MISS, UNK
			1	7118	BELT	USED
CLIMATE	ACCIDENT	WEATHER	-1	32147	CLEAR/CLOUD	CLEAR, CLOUDY, MISS, UNK
			1	7809	OTHER	RAIN, SNOW/ICE, OTHER
CURVE	ACCIDENT	ROAD-ALIG	-1	29224	STRAIGHT	STRAIGHT, MISS, UNK
			1	10732	CURVED	CURVED
DURBAN	ACCIDENT	LAND-USE	-1	22280	MISSING	MISSING, UNK
			1	17676	URB/RUR	URBAN, RURAL
HERR	DRIVER	DRIVER-ERR	-1	10463	NOERROR	NONE, MISS, UNK,
			1	29493	ERROR	SPEED,SIGN/SIGNAL,PASSING,ASLEEP,ETC.
PROFILE	ACCIDENT	ROAD-PRO	-1	32968	LEVEL	LEVEL, MISS, UNK
			1	6988	GRADE	GRADE
ROADLOC	ACCIDENT	IMP1LOC	-1	8624	ON ROAD	ON ROADWAY, MISS, UNK
			1	31332	OFF ROAD	ON SHOULDER, OFF ROADWAY
RURAL	ACCIDENT	LAND-USE	-1	32412	URBAN	URBAN, MISS, UNK
			1	7544	RURAL	RURAL
SEXY	DRIVER	SEX	-1	13065	FEMALE	FEMALE
			1	26891	MALE	MALE, MISS, UNK
STABLE	VEHICLE	PRESTAB	-1	34557	STABLE	TRACKING, NOT APPLICABLE, MISSING, UNK
			1	5399	NONSTABLE	SKIDDING, SPINNING, JACKKNIFING
STEER	VEHICLE	AVOID	-1	37142	NO AVOID	NO AVOIDANCE, MISS, UNK,
			1	2814	AVOID	AVOID VEHICLE, PEDESTRIAN, ETC
SURF	ACCIDENT	ROAD-SUR	-1	27848	DRY	DRY, MISS, UNK
			1	12108	ICY	WET, SNOW/ICE, OTHER
YOUTH	DRIVER	AGE	-1	18206	OLD	25 AND OVER
			1	21750	YOUNG	LESS THAN 25

	RANGE	MIN/MAX	MEAN	SD
SF	.56	1.01	1.38	.147
		1.57		
WB	37.9	83.5	102.81	7.3
		121.4		

2.3 COMPUTER TECHNIQUE

The general approach to computer analysis of the data was to:

1. Bring the data-set onto the active hard disk from the Managed Storage System in order to make the data available to the Central Processing Unit.

2. Select cases and variables from the original data-set and restructure variables into categories as required by the particular regression analysis.

3. Call the Biomedical Data Package (BMDP) and specify the logistic regression program desired for the analysis.

4. Specify the regression model to be used for the particular run.

5. Construct the prediction equation in a SAS program, using the coefficients of the variables provided by the logistic regression analysis, to predict probability of rollover for each accident. Compare the predicted probability of rollover with the actuality in the accident and compute the correlation coefficient over the 40 make/models. Sort the cases by make/model and obtain a plot of predicted versus actual probability of rollover by make/model.

6. Derive the value of r^2 for each model as explained in Section 3, Methods, of this report.

Steps 1, 2, and 3 use SAS and presuppose computer system Job Control Language (JCL) in order to direct the system to the appropriate libraries and data-sets.

3.0 METHODS

Previous studies of the relationship of the stability factor to the proportion of single vehicle accidents which result in rollover have addressed the problem at the make/model level. They estimate a linear model to relate the percent rollover with a make/model to its stability factor. In order to examine nonvehicle factors more exhaustively, it appears that an analysis at the accident level offers more precision. It is well known that logistic regression, a nonlinear procedure, offers advantages over linear regression when estimating a proportion based on individual observations where the proportion in question is represented by occurrence or nonoccurrence of the corresponding event. (See Cox[5] or Afifi and Clark[6] for more information.) In this case, the event is rollover while the basic observation is a single vehicle accident. Either a rollover occurs or it does not, so there is no question of the individual accident providing an estimate of the proportion of rollover.*

Nevertheless, a regression can estimate the proportion of rollover based on these single observations. Logistic regression is more powerful and accurate than simple linear regression in this context. One reason is that in linear regression the estimate of the proportion must inevitably go above one and below zero for some values of the independent variables. This distorts the process and makes linear regression less efficient and less accurate for this purpose. Logistic regression is, however, more difficult and costly to perform, since it is a nonlinear procedure. Fortunately, a convenient logistic regression package is available with BMDP which interfaces with SAS in the NIH computer. This enables logistic regression models to be constructed almost as easily as linear models. However, the cost of a logistic regression procedure on a very large data-set can be considerable; so that the runs to be performed must be chosen with some care.

The output of the logistic regression when run in BMDP gives various quantities of interest. The primary interest centers on the coefficients of the variables, particularly the coefficient of the variable of most interest. In this case, the stability factor will be of most interest while secondary interest will go to factors which may affect our estimate of the influence of the stability factor.

---

*In this report, a reference to "proportion of rollover" means as a proportion of all single vehicle accidents. Similarly, "probability of rollover" means a probability given a single vehicle accident and "rollover rate" likewise means a rate based on single vehicle accidents.

In addition to the coefficients themselves, their t-values (i.e., the ratios of the coefficient to their standard errors) are of most interest. As in an ordinary regression, the stability factor will be considered useful in predicting the proportion of rollovers if its coefficient is large, its t-value is large (the latter indicating high confidence) and if both remain large in the presence of other factors, (indicating that the influence of the stability factor is not due to the intervention of other factors with which the stability factor is associated).

The importance accruing to a coefficient due to its size can be judged by substituting various values for the variable and calculating the proportion implied by the logistic model. If the independent variables entering the logistic regression are X1, X2,...XN and the constant is C, then there is a linear function determined:

$$L = C + A1*X1 + A2*X2 + \ldots + AN*XN \qquad (1)$$

where A1 is the coefficient of X1, etc.

The predicted probability of rollover, P, according to the model determined by the logistic regression is

$$P = 1./(1. + EXP\ (-L)) \qquad (2)$$

Examples will be given in Section 4.3 which show how large the changes in P are which are induced by changes in one of the variables such as X1. The size of the effect is seen to be determined largely by the coefficient (A1 in the case of the variable X1). The quality of the logistic regression model in predicting rollover proportion will be judged partially by the parameters generated by the regression; these are the coefficients and their t-values.

The overall quality of fit of the logistic regression model at the accident level can also be judged in its performance in predicting the probability of rollover in an individual accident by the likelihood statistic for the model.

The BMDP logistic regression program shows the logarithm of the likelihood (log likelihood) for each model it produces. It is convenient to subtract from this value the log likelihood for the null model (with a constant only, no data variables). The result is a number which ranges from near zero for models with almost no predictive power to over 1900 for our model which performs best in predicting the probability of rollover on individual accidents. We called this measure the "likelihood

information statistic" (or LIS). (It bears some resemblance to the Kullback discrimination information statistic.) The LIS will provide the primary means of comparing the predictive capability of models at the accident level.

A model will also be judged by the goodness of its predictions of rollover proportions. These are of interest at the make/model level. Therefore, a second means of evaluating each logistic regression model is to project onto the make/model level and evaluate the agreement of predicted and actual rollover rates.

For this purpose, a series of SAS procedures (PROC FREQ, weighted PROC FREQ, MERGE, PROC CORR, PROC PLOT, etc.) were combined to achieve this projection and evaluation.

First, an actual proportion of rollover was computed for each make/model. Then a predicted proportion of rollover was calculated by summing over a make/model the value of P from Equation 2 for each single vehicle accident pertaining to that make/model and dividing the result by the number of such accidents.

The actual and predicted rollover rates were then compared two ways:

1. The Pearson product moment correlation coefficient, r, was calculated (using PROC CORR in SAS). The value r^2 is used for the comparisons.

2. The actual rates were plotted versus the predicted rates using PROC PLOT in SAS.

Thus, a high value of r^2 and a plot tightly clustered around a straight line shows a good fit for the model while a low value of r^2 and a plot in which there is a greater deal of vertical scatter from the best fitting line shows a poor fit. In addition, the values of predicted and actual proportions, should agree well, i.e., the best fitting line should be the one where "predicted" = "actual."

It should be remembered that this means of evaluating the model examines its properties as projected to the make/model level only. Any variable which does not change much from make/model to make/model is not really thoroughly evaluated (in this means) in its usefulness in predicting percent rollovers for individual accidents. Instead, all factors are evaluated in their ability to predict rollover tendency of make/models.

Fortunately, this is primarily what is wanted in evaluating the stability factor and hence this way of evaluating the model is particularly useful for the present purposes. Using this evaluation, a factor may show up as unimportant mainly because it

tends to be well balanced over make/models. If interest centers on a factor which does not change much from make/model to make/model (e.g., driver age), then it must be evaluated also by its coefficient and t-values in the logistic regression.

4.0 RESULTS

4.1 LOGISTIC REGRESSION

Table 5 presents the main numerical results of this study (Section 4.2 contains the important graphical results). Seven models, each resulting from a separate logistic regression, are represented in columns each headed by a model designation. The models designated are described in Table 6, where each short designation is followed by a brief description of the model and/or how it was produced. The first column of Table 5 lists all the variables which appear in any of the seven models as noted in the previous section (Table 4 shows the definition of the variables). The row corresponding to each variable gives at the intersection with each model's column, the coefficient of the variable in the model. The absolute t-value is also given in parentheses. Also given at the bottom of the first column are headings for two rows which give, respectively, the make/model based r^2 and the LIS for each model. In using this table, it is most useful to compare models with respect to various parameters.

First consider the Model 11F. It is the most complete model represented in Table 5. It includes more variables than any of the other six and its likelihood information statistic (1907) is the highest. Its make/model based r^2 (0.944) is not the highest but is not significantly different from the highest. Given that a positive coefficient means that higher values of a variable give a higher probability of rollover, all of the coefficients in Model 11F appear to have the correct sign (although in some cases there is not a strong a priori conception of which sign the coefficient should have). Appendix A explains how these 11 variables were chosen for this model. Appendix A also includes the results on an analysis containing 16 CARDfile variables, the most we could practically evaluate because of cost and time limitations. These included all variables that we considered most likely to correlate with rollover probability. The 11-variable model was only slightly less capable in its predictive capability (LIS = 1907 vs. 1942) and was much less costly to run. Appendix B examines the suitability of the logistic model.

Model 11F may be compared to the Model 11F-SF, the 10-factor model which is the same as the 11-factor base model except that it does not include the stability factor. The LIS and r^2 drop markedly from 1907 to 1478 and from 0.944 to 0.5272, respectively. This is indicative of the importance of the stability factor in predicting rollover. The effect on the LIS would have been larger except for the collinearity of the stability factor with the wheelbase variable (i.e., in general the wider a vehicle's track the longer its wheelbase). This

TABLE 5. LOGISTIC REGRESSION MODELS FOR ROLLOVER PROBABILITY CONDITIONAL ON SINGLE VEHICLE ACCIDENT*

MODEL FACTOR	SF ONLY	7 F	7F REDUCED	11 F	11F-SF	11F-SF-WB	11F-WB
1 SF	-4.903800	-3.965900	-4.483400	-4.090000	*****	*****	-4.934200
	(45.41)	(28.70)	(21.20)	(29.37)	*****	*****	(44.52)
2 WB	*****	-.029990	-.030302	-.028107	-.080502	*****	*****
	*****	(10.83)	(8.437)	(10.06)	(36.47)	*****	*****
3 RURAL	*****	.456340	.449650	.456570	.475520	.481750	.452140
	*****	(19.00)	(13.22)	(18.94)	(20.03)	(20.68)	(18.77)
4 DURBAN	*****	-.393490	-.443740	-.397440	-.392100	-.293670	-.369840
	*****	(17.13)	(13.44)	(17.03)	(17.02)	(13.08)	(15.99)
5 CURVE	*****	.26478	.25394	.26075	.242860	.254560	.26693
	*****	(15.34)	(10.45)	(15.00)	(14.20)	(15.20)	(15.37)
6 HERR	*****	.404720	.407270	.392950	.375640	.374080	.397090
	*****	(18.01)	(13.51)	(17.26)	(16.65)	(16.79)	(17.45)
7 STABLE	*****	.215000	.265380	.287710	.286430	.297370	.289080
	*****	(9.59)	(8.675)	(12.07)	(12.25)	(13.01)	(12.14)
8 YOUTH	*****	*****	*****	.066667	.018644	.026348	.085073
	*****	*****	*****	(4.02)	(1.146)	(1.652)	(5.15)
9 ALCAD	*****	*****	*****	.093045	.079714	.070307	.094859
	*****	*****	*****	(4.71)	(4.11)	(3.71)	(4.812)
10 BELT	*****	*****	*****	.094840	.084718	.109700	.101900
	*****	*****	*****	(4.63)	(4.20)	(5.56)	(4.983)
11 SURF	*****	*****	*****	-.170520	-.148740	-.142570	-.173010
	*****	*****	*****	(8.82)	(7.85)	(7.68)	(8.95)
CONSTANT	4.620500	6.598100	7.318400	6.664400	6.488000	-1.6256	4.9588
	(32.56)	(29.36)	(22.37)	(29.47)	(29.21)	(49.9)	(33.41)
r-squared (Make/ Model Based)	.907	.9448	.9467	.9444	.5272	.6812	.9296
LIS (See Text)	1109	1832	*****	1907	1478	781	1856

*See following page for notes on this table.

Notes pertaining to Table 5:

1. The models identified by the column headings are described in Table 6.

2. The factors represented by row headings are defined in Table 4.

3. The goodness of fit measures "make/model based r^2" and "LIS" are defined in the methodology section.

4. The entry corresponding to a given model and a given factor is a model coefficient (and an absolute t-value).

5. The entries in the last two rows are the goodness of fit measures (see Note 3).

TABLE 6. MODEL DESCRIPTION

Model	Description
SF ONLY	This results from a logistic regression based only on SF --stability factor.
7 F	This results from a logistic regression using the seven variables: SF, WB, RURAL, DURBAN, CURVE, HERR, STABLE.
11 F	This results from a logistic regression using the seven variables in 7F together with four others of lesser importance: YOUTH, ALCAD, BELT, SURF.
7F REDUCED	This results from a logistic regression using the seven variables in 7F but trying a reduced data-set of observations obtained by eliminating all make/models with over 1700 observations each. Seven make/models were eliminated leading to the exclusion of 22,000 observations to obtain the data-set for this logistic regression only.
11F-SF	This results from a logistic regression using all the variables in 11F except SF, the stability factor.
11F-SF-WB	This results from a logistic regression using all the variables in 11F except SF and WB, the wheelbase.
11F-WB	This results from a logistic regression using all the variables in 11F except WB.

allows WB to serve as a proxy for SF in this model. Some evidence in support of this statement is the Pearson correlation of SF with WB which is found to be 0.64, an appreciable value. Also, in support is the fact that the coefficient of WB goes from -0.0805 with SF out of the model to -0.0281 with SF in the model (i.e., a shrinkage by a factor of almost three). Further evidence of the tendency of WB to proxy for SF will be developed below. The very large drop in the make/model based r^2 is to be noted since that measure is actually less for this model (11F-SF) than for the smaller Model 11F-SF-WB, i.e., the current model without WB. This observation will be discussed in connection with the plots which shed some light on the reason for it.

Next, consider the Model 11F-WB, i.e., the 11-factor base model but without WB. Here, both the LIS and the make/model based r^2 are rather good (large) showing that not too much is lost in predictive power when WB is dropped while SF is kept (further indication that the power of WB in the Model 11F-SF was as a proxy for SF). However, WB is the strongest variable besides SF. Furthermore, notice that the change in the coefficient of SF on dropping WB from the model, i.e., from -4.090 to -4.934, is the largest change in this coefficient due to the addition or deletion of any group of variables. The change in the SF coefficient in dropping all other variables is only from -4.934 (at 11F-WB) to -4.904 (at SF only).

The 17% drop in magnitude of the coefficient of SF due to inclusion of WB (Model 11F) (from -4.934 to -4.090) while substantial is not large enough to threaten the conclusion that SF is a very strong predictor of rollover. This is particularly true when it is remembered that a proxy variable (collinear with the main predictor) lowers the coefficient of the main variable even without adding much to the total predictive power of the model. Thus, SF does very well without WB (r^2 = 0.9296, LIS = 1856) but WB does very poorly without SF (r^2 = 0.5272, LIS = 1478).

The presence of SF causes a threefold drop in the coefficient of WB while the presence of WB causes only a 17% drop in the coefficient of SF. Of course, all this is not to say that WB is not useful. WB used in conjunction with SF leads to significantly more accurate prediction problems than SF alone.

The Model 11F-SF-WB, i.e., the 11-factor base model with both SF and WB dropped has a predictably low LIS -- only 781. The make/model r^2 is 0.6812, much smaller than for any model containing SF but perhaps surprisingly large. This will be discussed further in connection with the plots.

Two models represented in Table 5 remain to be discussed -- the seven-factor basic model, 7F, and the same model fit on a reduced data-set, 7FR. The Model 7F is based on the seven factors in 11F

with the most combined predictive power (as evidenced by their t-values). They yield an LIS of 1832 (very respectable) and a make/model based r^2 of 0.9448 (essentially identical to that of 11F). The variables left out are of little ancillary use in predicting rollover.

The Model 7FR is the Model 7F, but fit to a database which excludes all make/models with over 1700 single vehicle accidents in the database in order to remove approximately 50% of the SVAs. This reduced data-set was produced because make/models with very high numbers of accidents may tend to dominate the logistic regression. In this manner, 33 make/models and 18,177 observations remained in an experiment to see if the make/models with fewer accidents yielded different estimates. The result was that the coefficient of SF actually increased a bit and all coefficients remained fairly close to their original values. It is concluded that there is no substantial bias when using the full data-set and that doing so does not exaggerate the significance of SF.

Further models are discussed in Appendix C where it is concluded that the urban/rural variable may be more important at the accident level than concluded here. Other conclusions are not affected.

4.2 GRAPHICAL RESULTS

Figures 1-6 show plots on a make/model basis of actual vs. predicted proportions of single vehicle accidents which are rollovers (Table 7 gives the make/model symbols and FARS codes). These plots show graphically the data on which the make/model r^2 values are based. For each make/model, the actual fraction of rollovers is plotted and compared to the predicted fraction based on the model (calculated by averaging the predicted fraction for each accident over the accidents for the given make/model). These plots show the predictive capability of each model at the make/model level. In this section, the plots are discussed individually.

TABLE 7. MAKE/MODELS CONSIDERED IN ROLLOVER STUDY

Make/Model	Symbol	Make/Model
3	A	Chevrolet Blazer S-10
4	B	Chevrolet Blazer
69	C	Chrysler Cordoba
107	D	AMC Concord
123	E	Ford Mustang
126	F	Ford LTD
201	G	Chevrolet Malibu
211	H	Oldsmobile Cutlass
215	I	Chevrolet Citation
225	J	Jeep CJ-5
227	K	Jeep CJ-7
271	L	Cherokee
607	M	Chrysler LeBaron
707	N	Dodge Diplomat
709	O	Dodge Miranda
1271	P	Ford Bronco
1403	Q	Mercury Capri
1406	R	Mercury Marquis
1801	S	Buick Century Regal
1815	T	Buick Skylark
1903	U	Cadillac DeVille/Brougham
2004	V	Chevrolet Corvette
2009	W	Chevrolet Camaro
2010	X	Chevrolet Monte Carlo
2113	Y	Chevrolet Chevette
2115	Z	Oldsmobile Omega
2201	a	Pontiac LeMans
2209	b	Pontiac Firebird
2215	c	Pontiac Phoenix
3032	d	VW Beetle
3036	e	VW Rabbit
3234	f	Audi 4000
3533	g	Datsun B210
3534	h	Datsun Z, ZX
3731	i	Honda Civic
4135	j	Mazda GLC
4631	k	Renault Le Car
4932	l	Toyota Corolla
4971	m	Toyota Landcruiser
8471	n	I H Scout

Figure 1 shows the plot for the 11-factor basic model. The agreement between actual and predicted rollover rate is striking. The symbol "B" (representing the make/model Chevy Blazer) is much further from the line of perfect agreement than any other. If this make/model did not have substantially more rollovers than predicted, the agreement would be even more striking.* The already high r^2 would be even larger.

Figure 2 shows the same plot for the seven-factor basic model. This plot is very similar to the previous one. The additional variables in the 11-factor model, therefore, had very little effect on the predictions at the make/model level.

Figure 3 shows the plot of actual vs. predicted rollover rate for the model generated by logistic regression using the variable set which is obtained when both WB and SF are dropped from the 11-variable basic set. As expected, the actual fraction of rollover for most make/models agrees very poorly with the corresponding predicted value. The r^2 value at 0.681 is deceptively high. This could be because the make/models particularly susceptible to rollover are somewhat overrepresented in rural accidents (Rural = 1) and in Texas accidents (DURBAN = -1). Because of the large sample sizes, this could lead to a small but systematic effect. However, notice that the total range in predicted rollover fraction is only from about 0.10 to about 0.16, while for the better models (7F and 11F) it ranges from 0.04 to 0.48. The actual fraction goes through a similar large change. To repeat: the agreement between actual and predicted rollover rates is worse here than in any other model represented in the plots of Figures 1-6.

Figure 4 shows the plot of actual vs. predicted for the model with only SF left out of the 11-factor model. The difference between the model represented here and that in the previous plot is that wheelbase is included. Since wheelbase is a relatively

*A satisfactory explanation of (at least the worst part of) this discrepancy may lie in the unusually small number of accidents representing the Blazer in this database. There were only 89 single vehicle accidents for the Blazer. Therefore, the observed rollover rate is not a very precise estimate of the expected. For comparison, only 10% of the make/models had fewer than 113 single vehicle accidents and the average make/model had nearly 1000. It is known that the 1983 Blazer data contained both 2-wheel drive and 4-wheel drive vehicles. These could not be differentiated in CARDfile and yet they have two different SFs. This ambiguity is also probably part of the reason for the discrepancy.

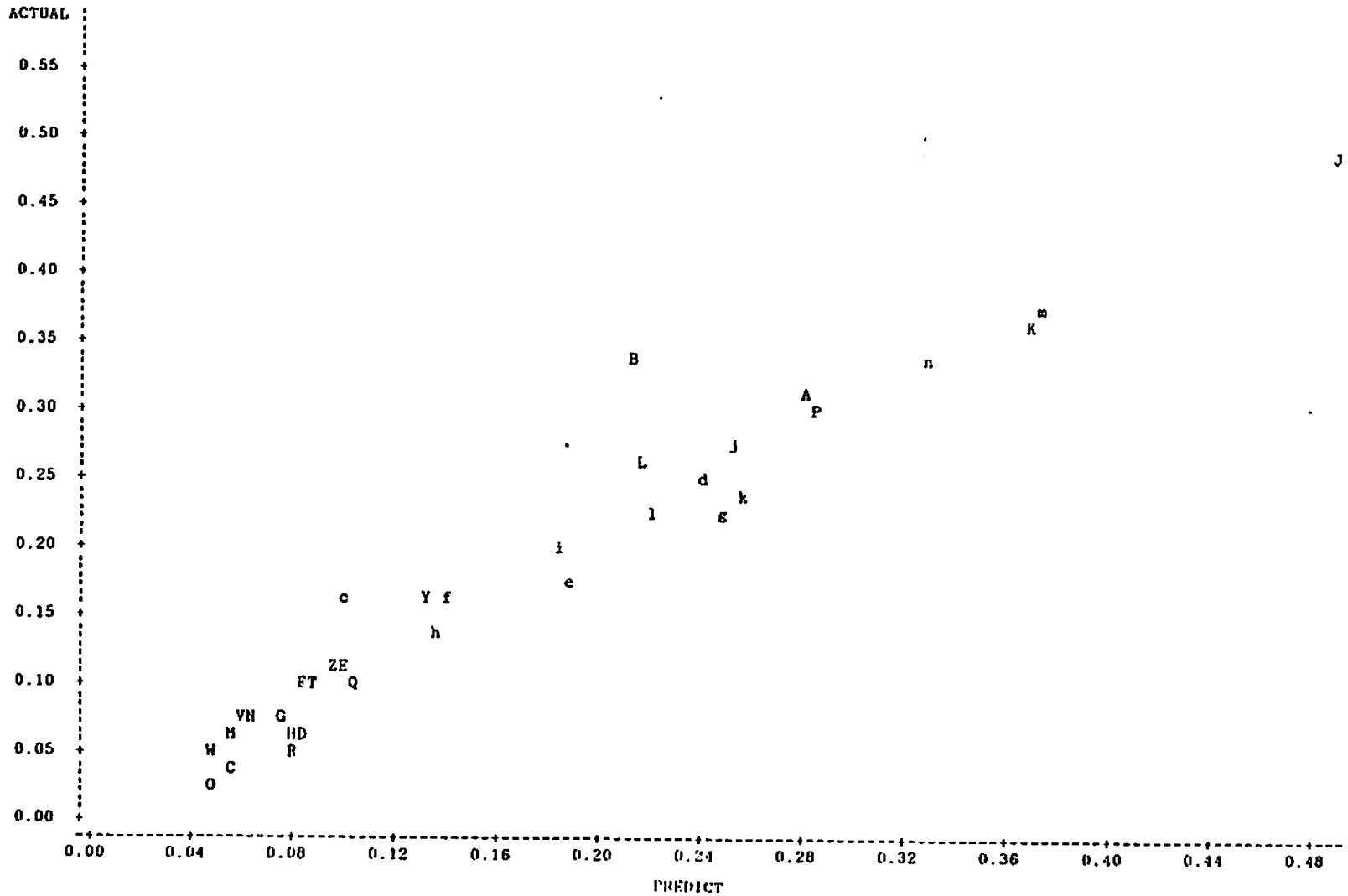


FIGURE 1. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR 11-FACTOR BASIC MODEL (11F)

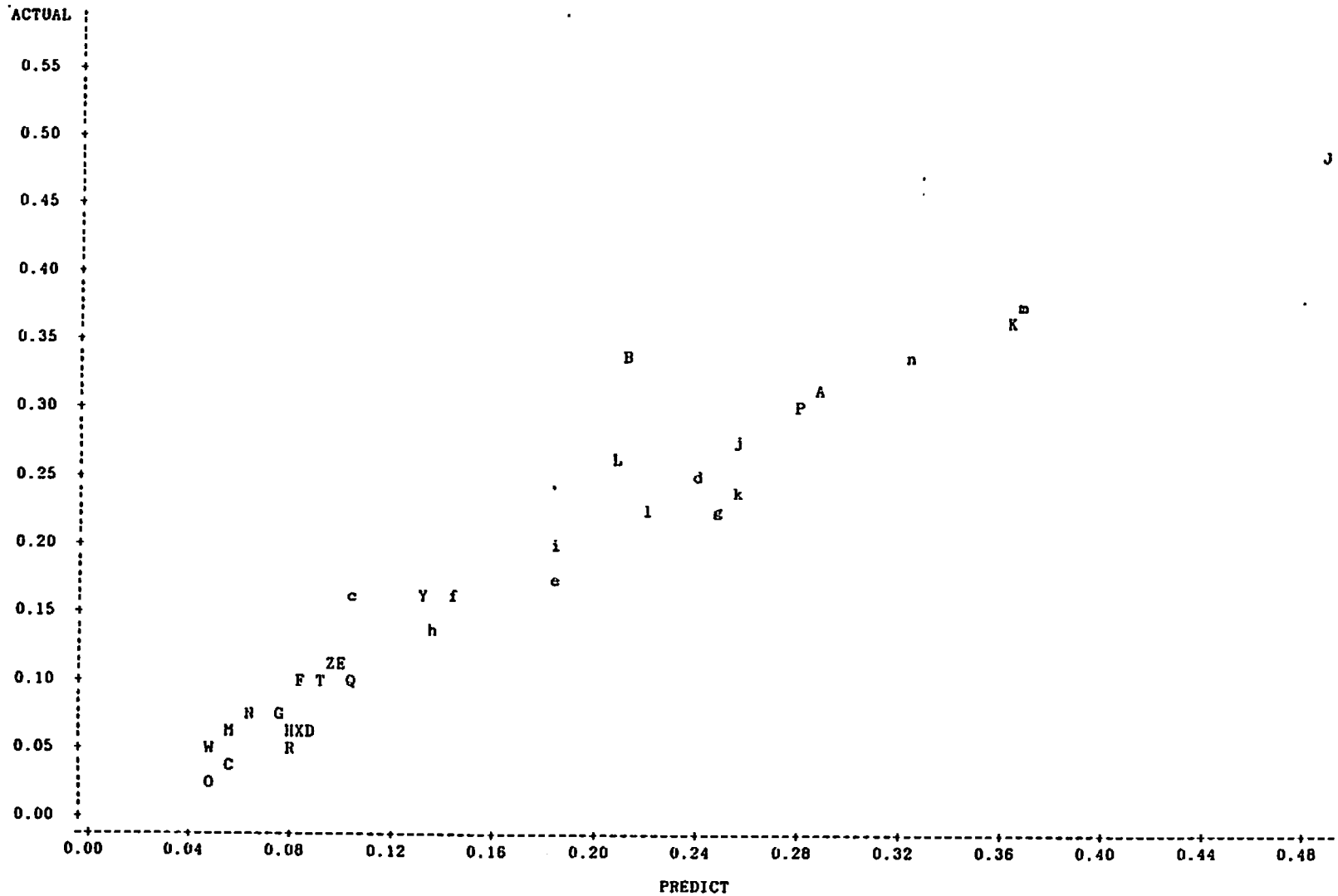


FIGURE 2. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR SEVEN-FACTOR BASIC MODEL (7F)

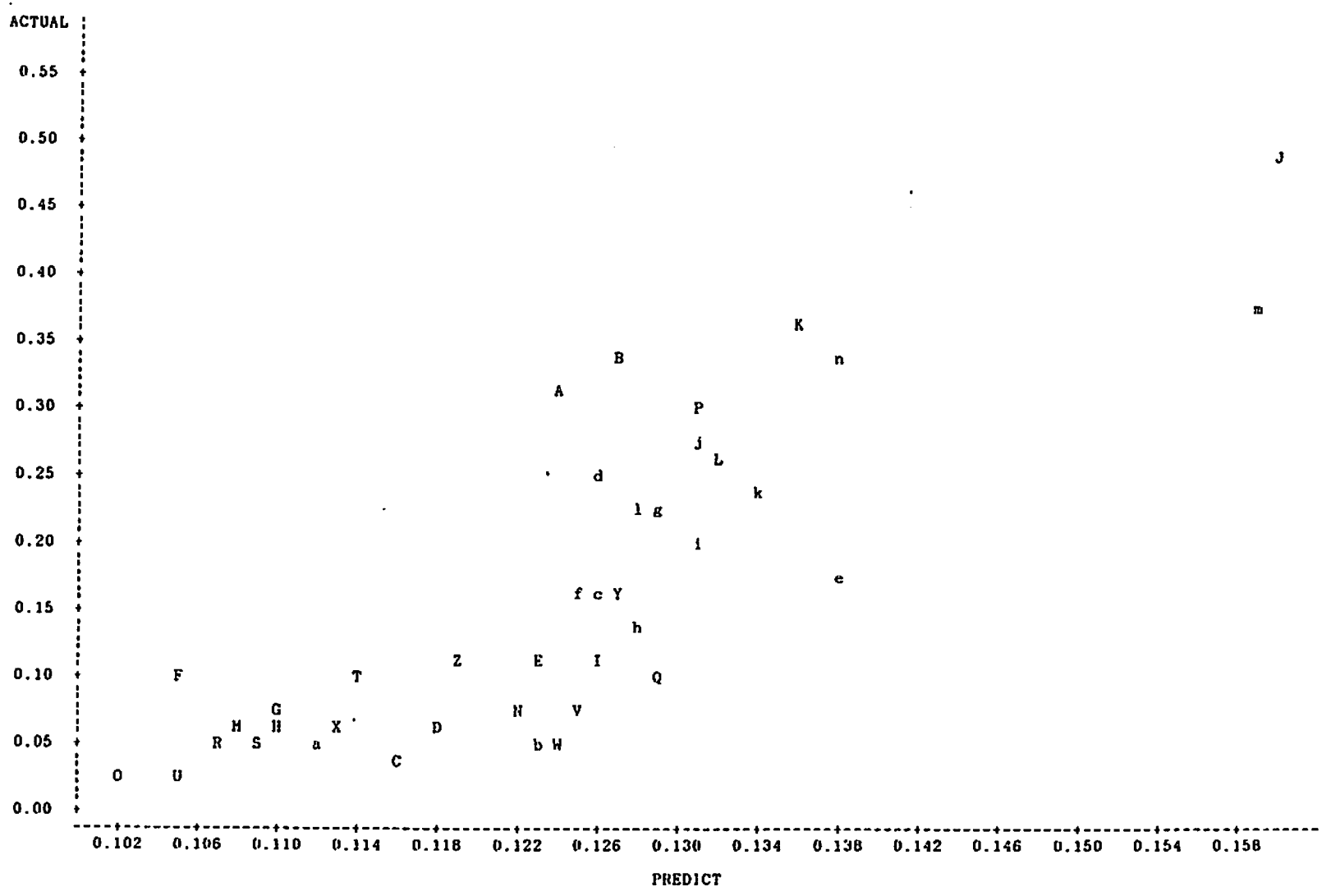


FIGURE 3. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR MODEL 11F-SF-WB

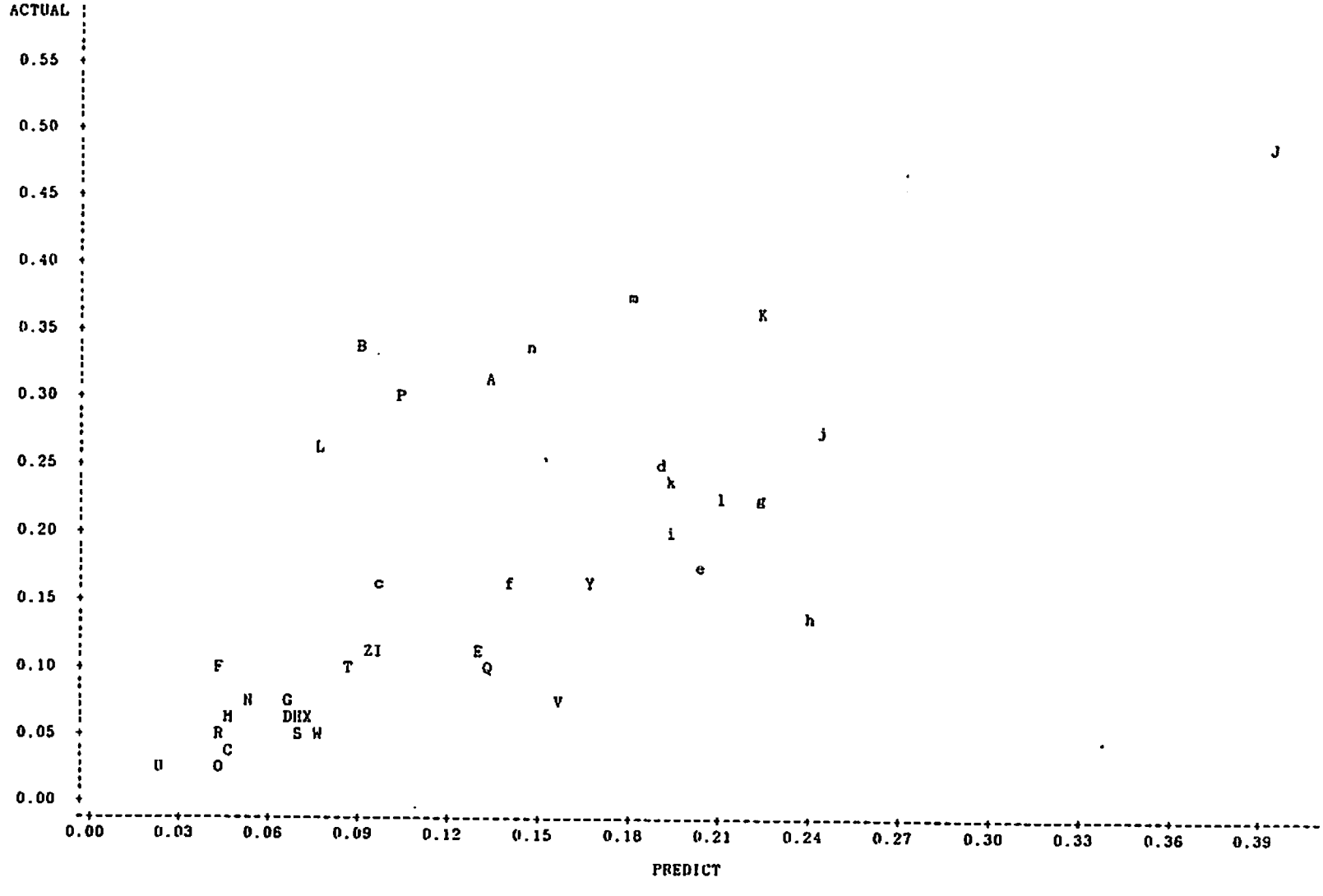


FIGURE 4. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR MODEL 11F-SF

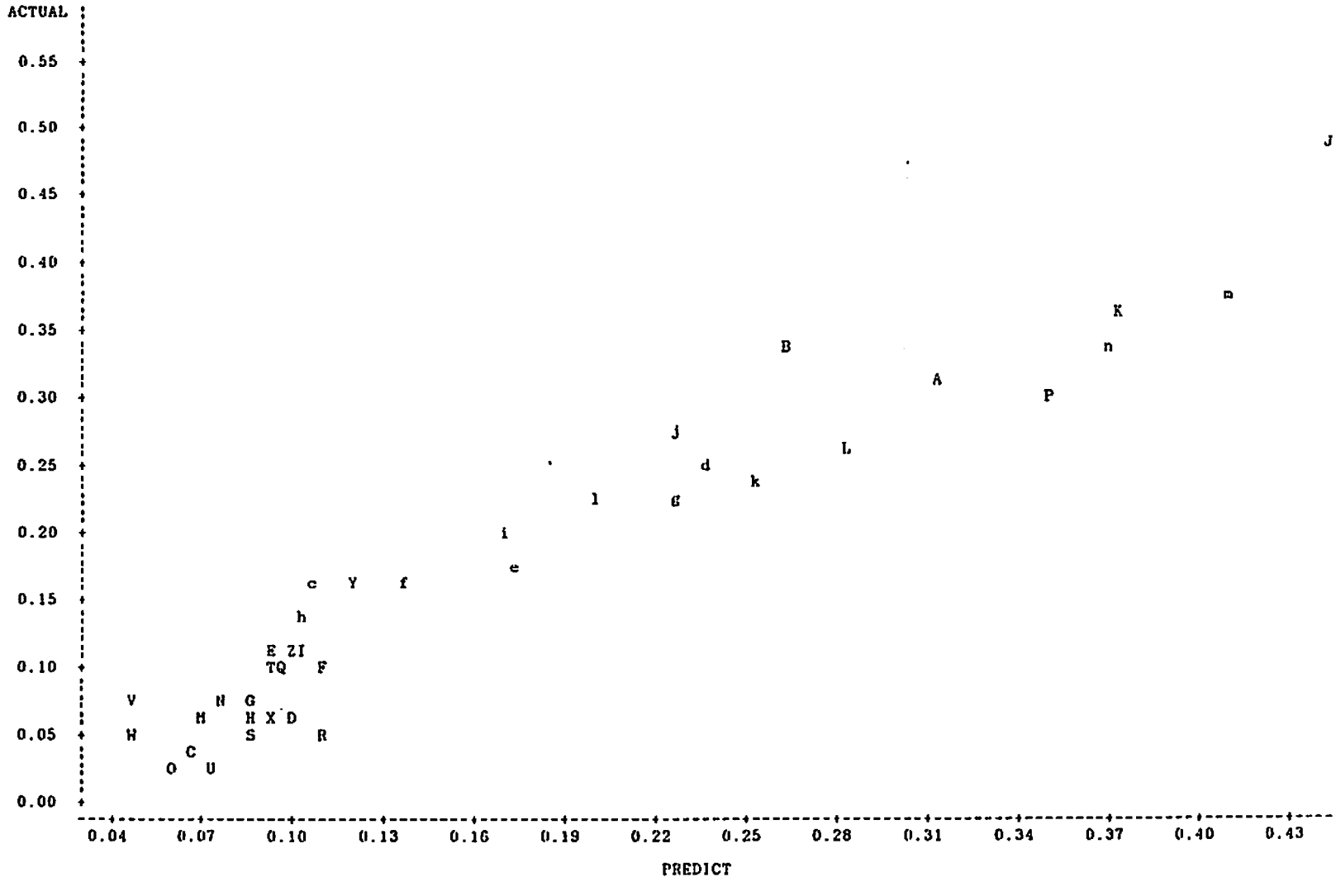


FIGURE 5. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR MODEL 11F-WB

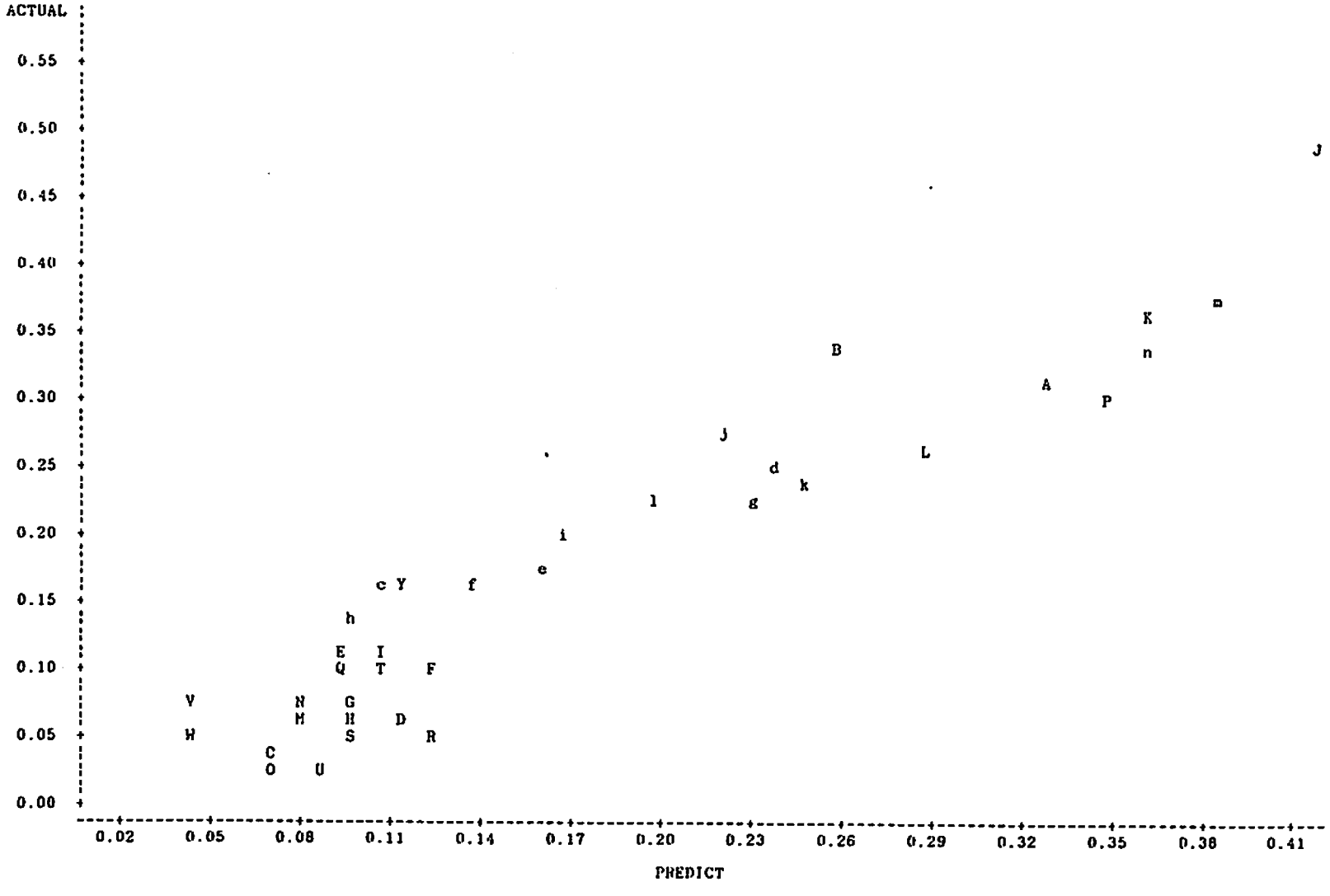


FIGURE 6. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR MODEL SF ONLY

strong variable (compared to all variables except SF), we would expect the agreement between actual and predicted to improve. It does substantially but the r^2 value actually decreases.* The agreement and r^2 would both be better except for a cluster of seven points labelled "L," "B," "P," "A," "n," "m," and "K." These points represent seven of the eight utility vehicles represented in our data. The eighth, "J," is a point by itself. It appears that based on a model with wheelbase and not SF the rollover rates of utility vehicles are being systematically underestimated. One possible explanation is that WB is acting as a proxy for SF here and in utility vehicles the relationship of WB to SF is different than in other vehicles (specifically lower SFs correspond to given WBs when considering utility vehicles vs. other vehicles; this has been verified with separate regressions). This may also be an effect of other vehicle geometry factors that were not tested, especially those relating to vehicle suspension characteristics.

Figure 5 shows the actual rollover fractions vs. the predicted for Model 11F-WB, i.e., the 11-factor model with the wheelbase left out. Although still quite good, the plot is noticeably different and the fit is somewhat worse than that for 11F or 7F.

Figure 6 is the final plot in this series. It shows actual vs. predicted for the model based on stability factor only. Although the agreement between actual and predicted is not as close as for other models involving SF, it is clearly much better than models not involving SF. The agreement is rather good. Other graphical results are found in Appendix C.

4.3 RELATIVE STRENGTH OF MODEL COEFFICIENTS WHEN APPLIED TO DATA

The question arises as to the strength and meaning of the coefficients in the logistic regression models. This section addresses this question by evaluating the rollover probability, P, for various values of the variables on which it depends (the inputs). Of interest is the variation in P as the inputs vary. A variety of examples will be examined. Values must be chosen for each variable and the value of P according to Equation 2 is calculated. P is then the estimated probability of a rollover given a single vehicle accident described by the chosen values of the variables. The seven-factor model is adequate to illustrate. (For the first illustration, three sets of values for each variable are chosen.)

*The r^2 value decreases while the agreement of predicted to actual increases substantially. There is no contradiction here since r^2 measures association, not agreement.

1. Choose the value of the stability factor to be its mean value over the data-set, i.e., 1.383. Also, choose the mean value for the wheelbase (102.867). For all the other variables (each is two level or dichotomous) choose the most frequent value. The result is P = 0.1084. This is to be compared with the mean overall rollover rate of 0.1229 which is satisfactorily close.

2. Choose the most frequent value for all dichotomous variables and the mean value for the wheelbase (both as before) but choose 1.1 for the stability factor. The result is P = 0.2714.

3. Choose the most frequent value for all dichotomous variables except HERR (driver error) and the mean for both WB and SF. The result is P = 0.0513.

The results show that changing the stability factor can have a large effect on predicted rollover rate but changing HERR (human error) can also. HERR has the second largest coefficient of the dichotomous, nonvehicle variables.

In order to further elucidate how influential the stability factor is when compared to the nonvehicle factors, the actual values seen in the data should be examined. Therefore, the value of P for each actual record was calculated in turn in three different ways corresponding to three cases:

Case 1: Choose SF and WB to be at their mean values (1.3825 and 102.8668 respectively) but let all other variables take their actual values.

Case 2: Choose each dichotomous variable to be at its overall most frequent value and choose WB to be at its mean value but choose SF to take its actual value.

Case 3: Choose all dichotomous variables at their most frequent value but let SF and WB both take their actual values.

Table 8 summarizes the nominal values for each variable and how the values of each variable are chosen for each case (Case 1, Case 2, Case 3).

In each case, a frequency distribution of P was taken over the data-set. The variability of P with respect to the variables taking their actual values was thus illustrated. The distribution of P for each case is shown in Tables 9, 10, and 11. The greatest variability is seen in Table 11 where both SF and WB take their actual values (Case 3). The second most variability is shown in Table 10 where only SF takes its actual values (Case

TABLE 8. NOMINAL VALUES FOR VARIABLES AND SPECIFICATIONS FOR CASES ONE, TWO, AND THREE

Variable	Nominal Values Mode/Mean	Case 1	Case 2	Case 3
SF	1.3821	nominal	actual	actual
WB	102.8076	nominal	nominal	actual
RURAL	-1	actual	nominal	nominal
DURBAN	-1	actual	nominal	nominal
CURVE	-1	actual	nominal	nominal
HERR	+1	actual	nominal	nominal
STABLE	-1	actual	nominal	nominal

TABLE 9. DISTRIBUTION OF p VALUES: CASE 1

PROB	FREQUENCY	CUM FREQ	PERCENT	CUM PERCENT
0.02420413	1644	1644	4.115	4.115
0.03673035	250	1894	0.626	4.740
0.04042004	265	2159	0.663	5.403
0.05167368	5956	8115	14.906	20.310
0.05278545	4000	12115	10.011	30.321
0.05819219	826	12941	2.067	32.388
0.06081534	147	13088	0.368	32.756
0.07729003	148	13236	0.370	33.126
0.07890708	1015	14251	2.540	35.667
0.08469584	588	14839	1.472	37.138
0.08645331	1859	16698	4.653	41.791
0.08674439	188	16886	0.471	42.261
0.09496287	268	17154	0.671	42.932
0.1090667	11496	28650	28.772	71.704
0.1218946	2328	30978	5.826	77.530
0.1245326	15	30993	0.038	77.568
0.127002	952	31945	2.383	79.950
0.138896	168	32113	0.420	80.371
0.1583829	417	32530	1.044	81.415
0.1721095	3581	36111	8.962	90.377
0.1758661	956	37067	2.393	92.770
0.1907645	1746	38813	4.370	97.139
0.2421827	79	38892	0.198	97.337
0.2659927	1064	39956	2.663	100.00

TABLE 10. DISTRIBUTION OF p VALUES: CASE 2

PROB	FREQUENCY	CUM FREQ	PERCENT	CUM PERCENT
0.05461683	10781	10781	26.982	26.982
0.07909856	577	11338	1.394	28.376
0.08821065	564	11901	1.412	29.788
0.09480199	161	1206	0.403	30.191
0.09826043	2456	14519	6.147	36.337
0.1018308	8918	23437	22.320	58.657
0.1093179	1974	25411	4.940	63.597
0.1172836	822	26233	2.057	65.655
0.1257477	1665	27898	4.167	69.822
0.1347295	120	28018	0.300	70.122
0.1543168	1686	29704	4.220	74.342
0.1595636	1654	31358	4.140	78.481
0.182002	1388	32746	3.474	81.955
0.2003889	269	33015	0.673	82.628
0.2068192	2433	35448	6.089	88.718
0.2134007	2404	37852	6.017	94.734
0.2201335	113	37965	0.283	95.017
0.2270175	89	38054	0.223	95.240
0.2485702	64	38118	0.160	95.400
0.2793644	246	38364	0.616	96.016
0.2956085	710	39074	1.77	97.793
0.3039329	505	39579	1.264	99.056
0.3209692	197	39776	0.493	99.550
0.3474325	180	39956	0.450	100.000

TABLE 11. DISTRIBUTION OF p VALUES: CASE 3

PROB	FREQUENCY	CUM FREQ	PERCENT	CUM PERCENT
0.04719182	10741	17041	26.882	26.882
0.05666985	161	10902	0.403	27.285
0.05794644	557	11459	1.394	28.679
0.06266175	40	11499	0.100	28.779
0.06719615	564	12063	1.412	30.191
0.0885887	6858	18921	17.164	47.355
0.09338594	1665	20586	4.167	51.522
0.1022607	549	21135	1.374	52.896
0.1035208	1974	23109	4.940	47.836
0.105013	2456	25565	6.147	63.983
0.1357015	273	25838	0.683	64.666
0.1382196	2060	27898	5.156	69.822
0.1458161	120	28018	0.300	70.122
0.1899682	1686	29704	4.220	74.342
0.1961461	1654	31358	4.140	78.481
0.2084673	89	31447	0.223	78.704
0.2173477	64	31511	0.160	78.864
0.2286528	1388	32899	3.474	82.338
0.2858292	2404	35303	6.017	88.355
0.2628943	269	35572	0.673	89.028
0.2632852	113	35685	0.283	89.311
0.2647723	2433	38118	6.089	95.400
0.2842909	710	38828	1.777	97.177
0.2938754	246	39074	0.616	97.793
0.3224219	257	39331	0.643	98.436
0.353574	197	39528	0.493	98.929
0.3663915	248	39776	0.621	99.550
0.487619	180	39956	0.450	100.000

2) and the least variability (but almost equal to Case 2) is seen in Table 9 (Case 1) where all the dichotomous variables take their actual values. These tables illustrate once again the preeminence of SF even at the accident level.

Table 12 shows all three cases summarized at the make/model level. Attention is called to the column for Case 1. The average value of P changes remarkably little from make/model to make/model. This is because stability factor and wheelbase are being held constant while the other factors take on their actual values. But when the P value is averaged over the make/model the nonvehicle factors balance out. This measures how "lucky" each make/model was in the mix of drivers and accident situations it was "dealt." It is seen that they all got remarkably even deals as far as the effect on P is concerned. In fact, all the P values in this column for Case 1 are contained in the interval 0.107 ± 0.017. This observation, in some measure, answers an objection that the effect of stability factor cannot be estimated because nonvehicle factors (driver and accident situation variables) are dominant. This suggests that <u>only</u> vehicle factors need to be known to fairly accurately estimate rollover rates based on these data.

TABLE 12. MEAN PROBABILITIES BY MAKE/MODEL FOR CASES 1, 2, AND 3

OBS	MAKE/ MODEL	FREQ	CASE 1	CASE 2	CASE 3
1	3	247	0.110683	0.279364	0.293875
2	4	90	0.114430	0.227017	0.208467
3	69	509	0.100008	0.079099	0.057946
4	107	540	0.091113	0.117284	0.102261
5	123	1846	0.104572	0.098260	0.105013
6	126	1461	0.095704	0.125748	0.093386
7	201	1171	0.094083	0.101831	0.088589
8	211	2282	0.097242	0.101831	0.088589
9	215	1197	0.105590	0.109318	0.103521
10	225	180	0.122022	0.347432	0.487619
11	227	249	0.115991	0.303933	0.366391
12	271	74	0.106848	0.248570	0.217348
13	607	388	0.092314	0.088211	0.067196
14	707	198	0.099177	0.088211	0.067196
15	709	49	0.090227	0.079099	0.057946
16	1271	710	0.112161	0.295609	0.284291
17	1403	587	0.106647	0.098260	0.105013
18	1406	204	0.094524	0.125748	0.093386
19	1801	1401	0.097423	0.101831	0.088589
20	1815	328	0.095022	0.109318	0.103521
21	1903	161	0.093769	0.094802	0.056670
22	2004	40	0.106508	0.054617	0.062662
23	2009	7165	0.105586	0.054617	0.047192
24	2010	1659	0.100324	0.101831	0.088589
25	2013	273	0.107246	0.117284	0.135701
26	2115	182	0.099582	0.109318	0.103521
27	2201	355	0.097102	0.101831	0.088589
28	2209	3585	0.104561	0.054617	0.047192
29	2215	267	0.107366	0.109318	0.103521
30	3032	2404	0.105237	0.213401	0.258529
31	3036	1686	0.109095	0.154317	0.189968
32	3234	120	0.107476	0.134729	0.145816
33	3533	2433	0.105796	0.206819	0.264772
34	3534	2060	0.108007	0.101831	0.138220
35	3831	1654	0.104378	0.159564	0.296146
36	4135	260	0.109780	0.200389	0.262894
37	4631	113	0.109349	0.220134	0.263285
38	4932	1388	0.106872	0.182002	0.228653
39	4971	197	0.124003	0.320969	0.353574
40	8471	257	0.114938	0.303933	0.322422

5.0 CONCLUSIONS

The stability factor is by far the most important variable among those we examined for predicting rollover rate. At the accident level using the full data-set, stability factor was the most important single factor but other factors especially HERR (driver error) and RURAL (whether the accident took place in an urban or rural location) were also quite important.* Since the variables were chosen to be those most likely to predict rollover rate, the evidence is that the stability factor is strongest among those variables available when related to accident data. Only wheelbase had an appreciable effect on the coefficient of the stability factor when it was in the regression. Nonvehicle factors seem to be of practically no consequence in determining the strength and form of the relationship between rollover and the stability factor. Other vehicle factors relating to the suspension, tires, etc., may also be of importance but were not evaluated for this study.

As important as the stability factor was in the accident level logistic regression, its dominance was magnified when the predictors were examined at the make/model level. When the predicted logistic model rollover rates were aggregated to the make/model level and compared with actual rates, it was found that a model with only stability factor showed an r^2 of 0.907 while a model based on all other factors except stability factor had an r^2 of only 0.527. When the plots of actual vs. predicted rollover rates at the make/model level were examined, a striking ability to predict was seen for the seven- and eleven-factor models. Even the stability factor alone could in most cases predict rollover rate fairly accurately. Without the stability factor, the agreement between predicted and actual was very poor.

At the make/model level even the 11-factor model is dominated by the stability factor. The other variables provide corrections to produce a predictor which is very accurate in predicting the rollover rate of each make/model (except the Chevy Blazer whose rollover rate is underpredicted by 37%). The previous results of Kelley/Robertson and Harwin/Brewer in finding the stability factor important for predicting rollover rate have been confirmed and strengthened by these results. Those studies were strictly at the make/model level. We have, in addition, shown that

*When the reduced data-set with Texas left out is used, the variable RURAL (distinguishing urban from rural) becomes quite important at the accident level but still of negligible importance at the make/model level (see Appendix C).

stability factor is the single most important factor at the accident level and the strength of its effect is almost unaffected by nonvehicle factors. The fact that wheelbase produces a small but significant change in the strength of the stability factor effect (17%) is no doubt due to a collinearity between the two factors. The stability factor is clearly far more useful than the wheelbase in predicting rollover and its presence in the model produces a very large change in the wheelbase coefficient (a threefold shrinkage). Nevertheless, a model with both stability factor and wheelbase predicts rollover significantly better than stability factor alone. Perhaps other vehicle factors not considered here would also work well with stability factor in predicting the percent of single vehicle accidents which are rollover (or the probability of rollover given a single vehicle accident).

It was shown that although nonvehicle factors are of importance at the accident level, the nonvehicle factors studied here had very little influence on the predicted rollover rates at the make/model level. A balancing of these factors over make/models greatly attenuated any effect they have at the accident level.

Because over half the accident records had missing values for the variable RURAL, a special test was made and reported on in Appendix C to further examine the importance of this variable in rollover. It is concluded in Appendix C that RURAL is as important a variable in predicting rollover at the accident level, perhaps even more important than stability factor. However, at the make/model level it is of little use. The overriding importance of stability factor at the make/model level is confirmed in Appendix C.

It must be recognized that CARDfile represents driver and environmental factors by a limited number of variables. There could, for example, be a driver factor which was inadequately represented (or "captured") by CARDfile variables yet which was very influential on the probability of a single vehicle accident being a rollover. It could also be the case that this factor was not evenly distributed over make/models with the result that what appeared to be vehicle effects were actually evidence of the action of this unseen variable on the fraction of rollover. This hypothesis is unlikely. The following observations concerning the data and results given here go against this hypothesis.

1. The driver and environmental variables which were represented in CARDfile were not very influential at all at the make/model level. This was due to a combination of the strength of SF and WB even at the accident level and of the balancing of nonvehicle factors upon aggregation to the make/model level.

2. There is not much variance left for a new factor to explain as the factors currently represented in the database now explain 95% of the variance of probability of rollover (in the set of single vehicle accidents).

3. A factor which would "steal" the predictive power of stability factor would have to be highly correlated with it. This is not in itself in any way impossible but this puts another condition on an already strained hypothesis.

APPENDIX A

SOME INITIAL VARIABLE SELECTION

Five variables were considered in this study but not included in the Model 11F. These variables are: (refer to Table 4 for definitions)

1. SEXY
2. CLIMATE
3. PROFILE
4. STEER
5. ROADLOC

The variables SEXY and CLIMATE were particularly weak. They were rejected by stepwise procedures early on in the study. For example with SF, YOUTH, RURAL, ALCAD, CURVE, SURF, and PROFILE in the model SEXY had an F to enter of 0.51 and CLIMATE an F to enter of 0.0 (CLIMATE is highly collinear with SURF). In other words, both variables would have t-values less than one if added.

As for PROFILE, STEER, and ROADLOC they had t-values relatively small (compared to the variables retained in 11F) in some larger regressions. A 10-variable regression was run with the following variables:

Variable		t-Value
1.	SF	(29.1)
2.	WB	(10.4)
3.	RURAL	(20.0)
4.	DURBAN	(13.2)
5.	CURVE	(18.3)
6.	YOUTH	(6.2)
7.	ALCAD	(6.5)
8.	BELT	(5.3)

9. PROFILE (0.52)

10. STEER (1.65)

The t-values are indicated (in absolute value) after each variable. PROFILE and STEER were dropped because the t-values were too small, (0.52 and 1.65, respectively). In another 10-variable regression, ROADLOC had a t-value of 2.94. It was decided to drop ROADLOC in view of the very large sample size and the considerably larger t-values for the variables that were kept.

The 11 basic variables were chosen in this manner and the variables were used in the largest model considered in the results section (Model 11F) and this model provided the basis for various tests. The somewhat reduced model, 7F, formed the basis of the other tests. It was decided, however, to run all 16 factors in one logistic regression to provide a check on the relative importance of the variables. Table A-1 shows the results of this regression. Observe the following:

1. The make/model r^2 (0.9455) has increased only negligibly over those for Models 11F and 7F (0.9444).

2. The LIS (at 1942) is only slightly larger than that for Model 11F (1907).

3. The coefficients of all variables in Model 11F change only slightly in the 16-variable model.

4. The variables SEXY, CLIMATE, and PROFILE have negligible t-values.

5. The variables STEER and ROADLOC have appreciable t-values. The t-value for STEER 5.402 has actually increased over its value in a smaller (included) regression.

Although STEER now appears to be as strong as some of the weaker variables in the Model 11F, STEER and ROADLOC are clearly of marginal importance and their being left out of 11F is deemed to be of no consequences.

TABLE A-1. SIXTEEN VARIABLE MODEL

VARIABLE	COEFFICIENT	t-value
SEXY	0.013408	0.7957
YOUTH	0.061619	3.692
RURAL	0.45076	18.62
STEER	0.17672	5.4
ALCAD	0.088506	4.39
CURVE	0.26051	14.72
SURF	-0.18826	7.18
CLIMATE	0.024405	0.853
PROFILE	0.033899	1.415
DURBAN	-0.37045	14.75
BELT	0.093943	4.58
ROADLOC	0.079039	3.14
STABLE	0.29229	12.06
HERR	0.39158	16.19
WB	-0.027446	9.8
SF	-4.1263	29.58
CONSTANT	6.7286	29.36

MAKE/MODEL BASED R-SQD=.9455

LIS=1942

APPENDIX B

EXAMINATION OF THE LOGISTIC MODEL

A question was raised concerning the adequacy of the logistic model to represent probability of rollover given a single vehicle accident (conditional on SF and the other variables). The answer lies partially in the quality of fit of the Models 7F and 11F demonstrated by plots at the make/model level and discussed in Section 4.2.

It is always possible that using a different specification of the model form (i.e., different from the logistic function) would improve the quality of the fit. To examine possible consequences of such a fit, a logistic regression using the seven basic factors was run but included also were quadratic and cubic terms -- one in the square and the other in the cube of the linear function from the basic seven-factor model. Thus, if the probability of rollover according to the seven-factor model is

$$P = 1./(1. + EXP\ (-L7))$$

then the new model was based on the same seven factors plus L7 x L7 and L7 x L7 x L7.

This model should capture any small deviation from the logit function which would be useful.

The LIS increased from 1837 for the basic seven-factor model to 1851 for the model including quadratic and cubic terms. This was less than a third of the change in LIS in going from 7F to 11F (LIS = 1907), i.e., due to the addition of the four least important variables. A plot of the predicted rollover rate for the two models (7F and 7F + cubic) is plotted in Figure B-1. The two models agree almost exactly in which make/models have higher rates than others but a slight curve in the plot can be seen. It is concluded that the basic conclusions drawn from logistic regression using the straight logit model would probably be unaffected by using a more suitable probability function. However, the probabilities predicted could change somewhat, (probably by 20% or less).

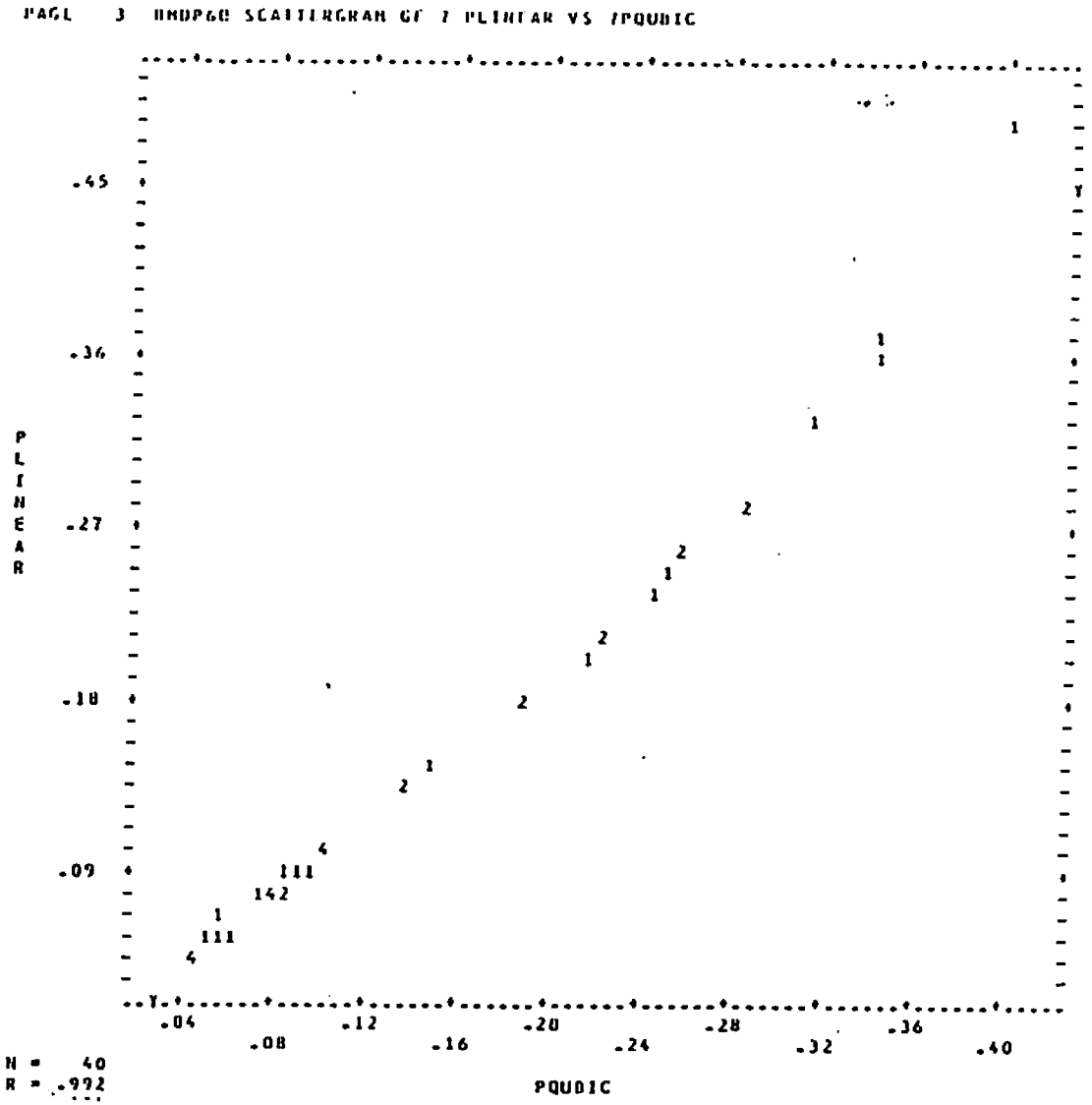


FIGURE B-1. SCATTERGRAM OF 7P LINEAR VERSUS 7P-CUBIC

APPENDIX C

FURTHER EXPLORATIONS OF THE URBAN-RURAL VARIABLE

In the main body of this report, the regressions are unable to develop the full usefulness of the variable RURAL (whether the accident took place in an urban or rural location) because 56% of the accident records had missing values for this variable (these were the Texas accidents). It was decided (as in Harwin and Brewer[1]) to do a special analysis of the accidents from Maryland and Washington which all have good values for the variable RURAL. Table C-1 shows the numerical results of that analysis. Three logistic regressions were run on the data for Maryland and Washington only. First, the seven-factor model (without the variable DURBAN) was run on these data. The variable DURBAN was dropped because it takes the same value, +1, over all the accident data from these two states. The column headed "All 6 Variables" gives results for this regression. Note that the t-value of RURAL exceeds that of SF in this regression showing some evidence that RURAL is an important variable in the regression. The product of the coefficient with the standard deviation, the "Beta value," for the variable SF is 0.497 while for the variable RURAL it is 0.457. This would suggest that the variables SF and RURAL are comparable in their influence. A more decisive test is to see what happens to the LIS (or the log likelihood) when the variable is dropped from the equation. The LIS values in this table are not to be compared with those in Table 5 since the data-set is quite different. However, it is very instructive to compare the LIS values of the three runs represented in Table C-1. The second and third runs represented there are for the models when first RURAL alone and then SF alone is dropped from the six-factor model. The LIS value falls more when RURAL is dropped (with SF left in) than when SF is dropped (with RURAL left in). The decrease in LIS value on dropping RURAL is almost twice the decrease on dropping SF (note that decreases in LIS are the same as decreases in the log likelihood since the LIS differs by a constant from the log likelihood).

In summary, it appears that the variable RURAL is as strong a predictor as SF or perhaps stronger in predicting rollover at the accident level. Of greater importance in the evaluation of SF as a predictor of rollover is the fact that, even in this context, the coefficient of SF changes only by a small amount (from -3.53 to -3.28) when RURAL is added to the regression.

In view of the results at the make/model level in the main body of this report, it is not to be expected that the strength of the variable RURAL would hold up at the make/model level.

TABLE C-1. LOGISTIC REGRESSION MODELS FOR URBAN/RURAL ANALYSIS

	Model		
Variables	All 6 Factors	6F less RURAL	6F less SF
SF	-3.2807 (15.31)	-3.532 (16.7)	*****
WB	-.035268 (7.895)	-.032057 (7.279)	-.08240 (24.67)
RURAL	.46226 (19.32)	******	.48182 (20.31)
CURVE	.16014 (6.742)	.19381 (8.283)	.14492 (6.161)
HERR	.48666 (13.08)	.49140 (13.32)	.47633 (12.84
STABLE	.25376 (10.38)	.28475 (11.83)	.25480 (10.53)
CONSTANT	5.7796 (17.29)	5.8312 (17.70)	6.1364 (18.61)
LIS	944	750	828
Make/Model r-squared	.9300	.9185	.6211

All three models represented in Table C-1 were tested at the make/model level and compared both in regard to their make/model based r^2 values (given in Table C-1) and in regard to plots of predicted and average rollover rates as seen in Figures C-1, C-2, and C-3. As can be seen, the predictive power at the make/model level suffers relatively little when RURAL is dropped from the equation but the predictive power is greatly reduced when SF is dropped. This is consistent with other results in this report.

To summarize, the results of the study of the relative importance of the variable RURAL when evaluated on those states for which the variable is not missing:

1. RURAL is an important variable in predicting rollover at the accident level; possibly the strongest single variable examined. However, its presence or absence has little effect on the coefficient of SF.

2. Its strength in predicting rollover at the make/model level is greatly reduced; at this level RURAL is not nearly as important as SF.

The reason why the variable RURAL loses so much predictive value in passing from the accident level to the make/model level is as demonstrated in this report - there is, in general, a balancing of nonvehicle factors over make/models so that such factors are of little importance in models for predicting rollover once the results are aggregated to the make/model level.

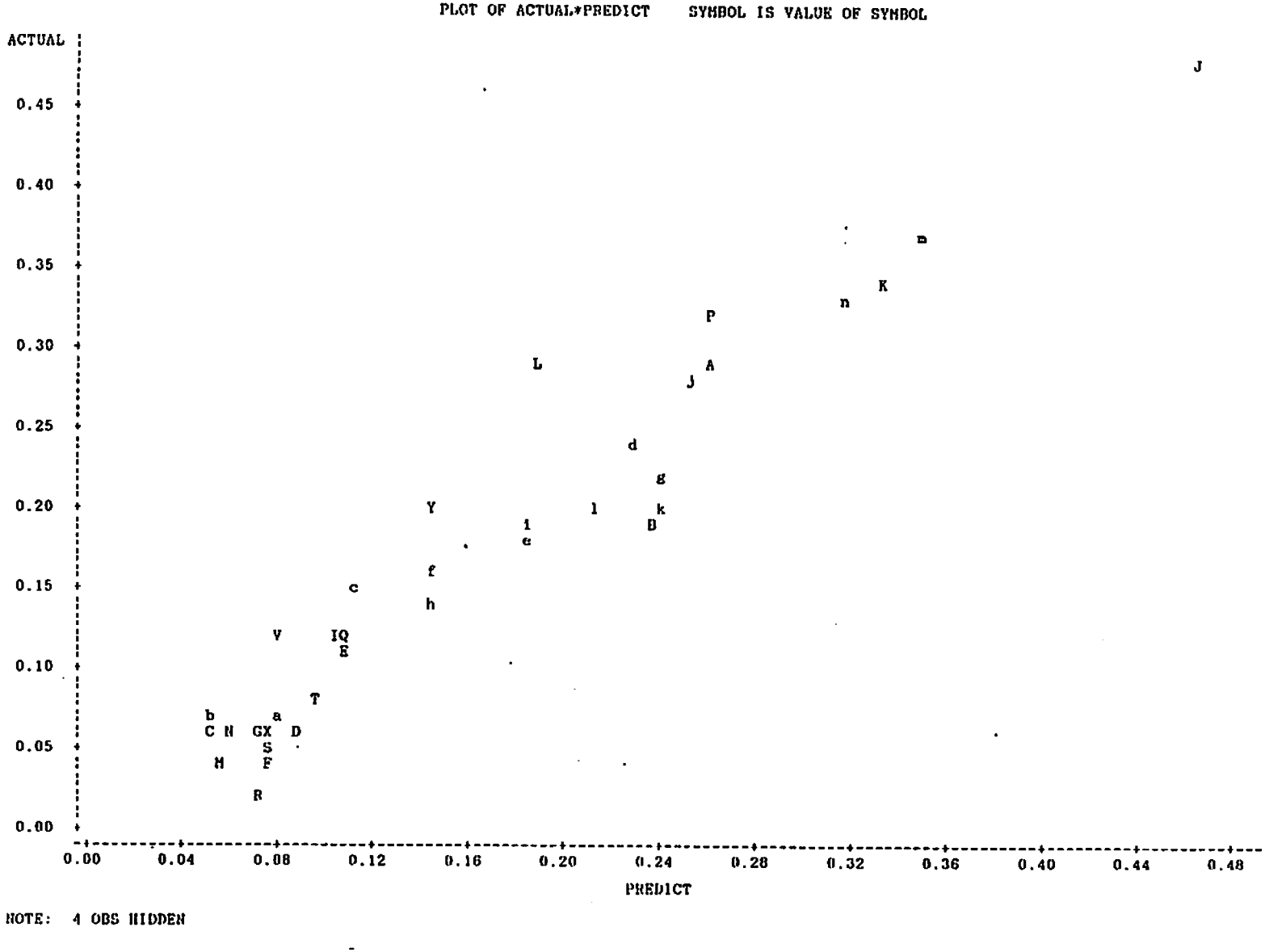


FIGURE C-1. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR MODEL 6F

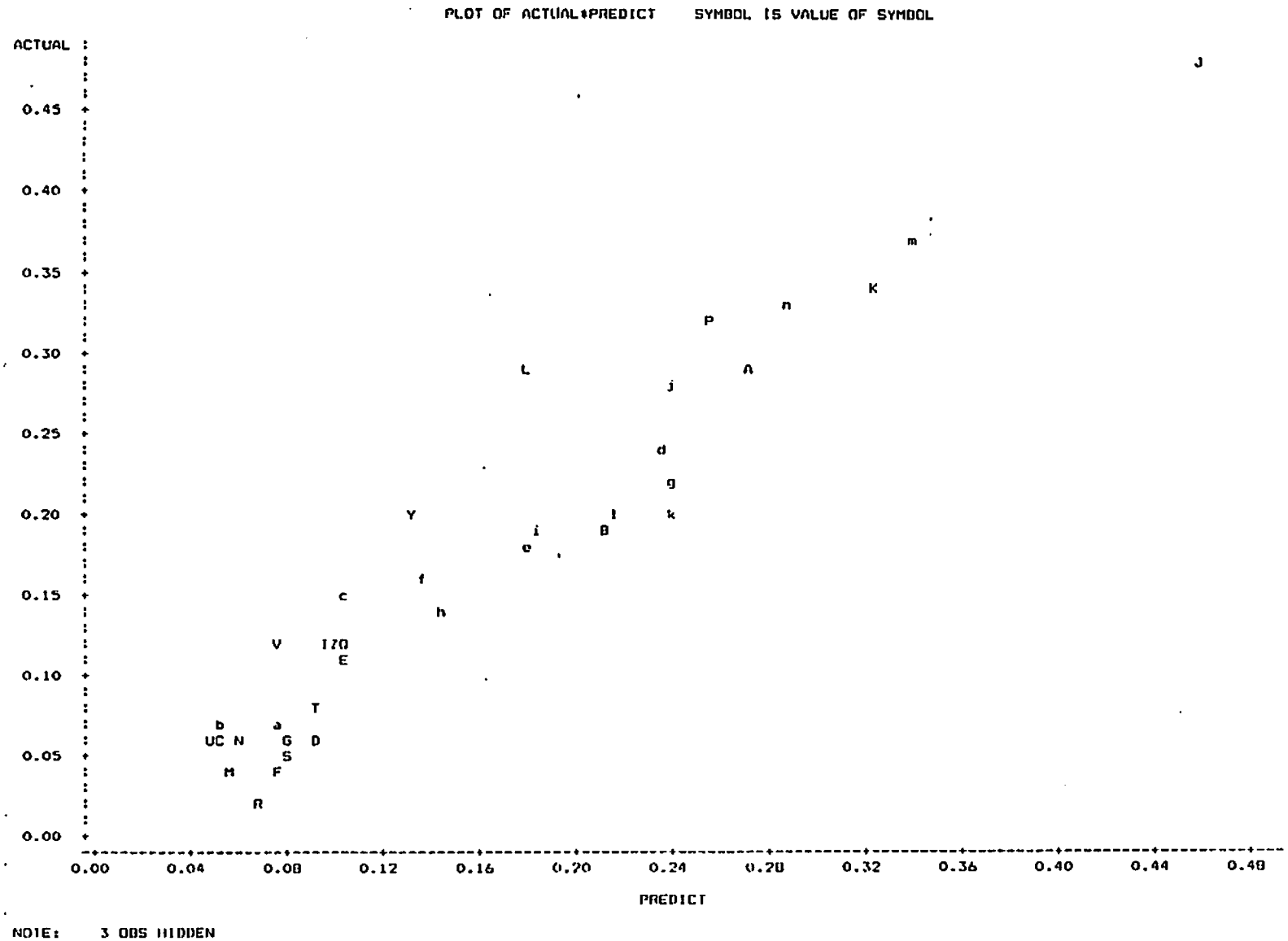


FIGURE C-2. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR MODEL 6F - RURAL

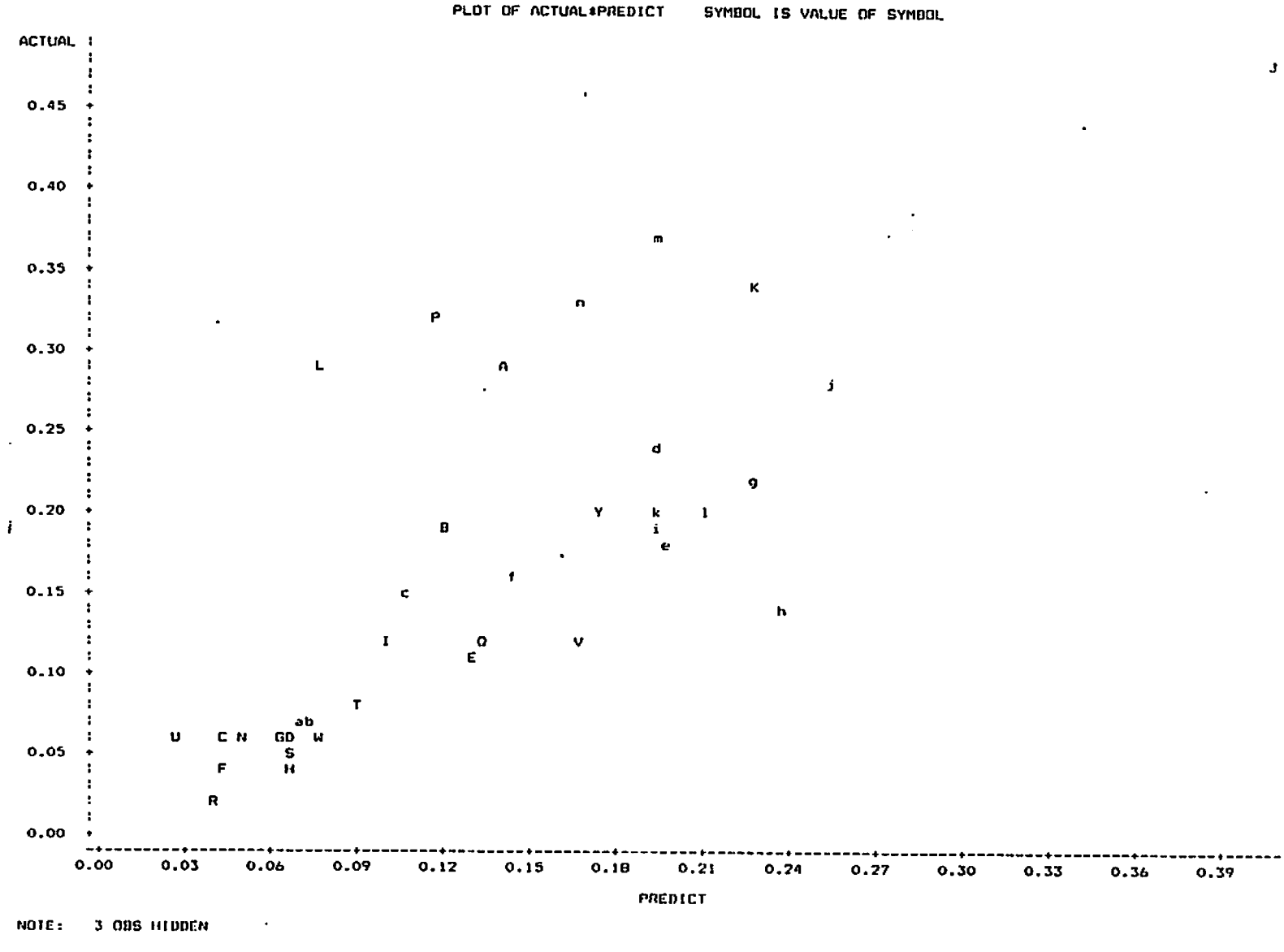


FIGURE C-3. ACTUAL VERSUS PREDICTED ROLLOVER RATES FOR MODEL 6F - SF

REFERENCES

1. Robertson, L. S. and A. B. Kelley, The Role of Stability in Rollover - Initiated Fatal Motor Vehicle Crashes Under On-Road Driving Conditions, Automotive Exchange Group, May 1986.

2. Harwin, E. A. and H. K. Brewer, Analysis of the Relationship Between Vehicle Rollover Stability and Rollover Risk Using the NHTSA CARDfile Accident Database, National Highway Traffic Safety Administration, 1987.

3. Salvatore, S., P. Mengert, and R. Walter, CARDfile Data Base Representativeness Phase I: General Characteristics Including Populations, Vehicles, Roads, and Fatal Accidents, Transportation Systems Center, Cambridge, MA, DOT-TSC-HS802-PM-88-16, August 1988.

4. Edwards, M., A Database for Crash Avoidance Research, SAE Special Publication No. 699, 1987.

5. Cox, D. R., The Analysis of Binary Data, Methuen, London, 1970.

6. Afifi, A. A., and Virginia Clark, Computer-Aided Multivariate Analysis, Lifetime Learning Publications, Belmont, CA, 1984.

7. Garrott, W. R., M. W. Monk, and J. P. Chrstos, Vehicle Inertial Parameters-Measured Values and Approximations, SAE Technical Paper Series, 881767.

U.S. Department
of Transportation

**Research and
Special Programs
Administration**

Kendall Square
Cambridge, Massachusetts 02142

Official Business
Penalty for Private Use $300

Postage and Fees Paid
Research and Special
Programs Administration
DOT 513