STATISTICAL ESTIMATION OF ROLLOVER RISK

Peter Mengert
Santo Salvatore
Robert DiSario
Robert Walter

Review Copy
March 10, 1988

Prepared for
National Highway Traffic Safety Administration
Office of Crash
Avoidance Research

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

PROBLEM

Studies have indicated that vehicles with a high center of gravity are more likely to roll over when involved in single vehicle accidents (reference 1 and 2). The popularity of the utility vehicle has prompted both Congressional and NHTSA action to further examine this issue. Utility vehicles, with their high centers of gravity and increased road clearance, are designed for both on- and off-road use. Most of the studies 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 and by Harwin and Brewer, 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. During NHTSA review of these results, it was 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.

---

*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.

APPROACH

The 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 CARDfile database which contains the police accident reports from six states - Texas (TX), Maryland (MD), Washington (WA), Indiana (IN), Michigan (MI), and Pennsylvania (PA). From over 2 million accidents in TX, MD, and WA, (1983 to 1985) 39,956 single vehicle accidents (SVAs) were analyzed. This analysis used 40 different make/models of utility vehicles, and domestic and imported passenger cars with known stability and other vehicle factors (see Table E-1; note that Tables E-1, E-2, and E-3 are identical to Tables 3, 4, and 5 respectively in the body of this report). In the data-set, the stability factor varied from 1.01 to 1.57. The data contained 4910 rollovers (RO). 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 E-2. During a preliminary analysis, we identified those variables that were most highly correlated with rollover. 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 E-3. These results show that at the accident level the variables which are very useful in

Table E-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.	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 E-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 E-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 (45.41)	-3.965900 (28.70)	-4.483400 (21.20)	-4.090000 (29.37)	***** *****	***** *****	-4.934200 (44.52)
2 WB	***** *****	-.029990 (10.83)	-.030302 (8.437)	-.028107 (10.06)	-.080502 (36.47)	***** *****	***** *****
3 RURAL	***** *****	.456340 (19.00)	.449650 (13.22)	.456570 (18.94)	.475520 (20.03)	.481750 (20.68)	.452140 (18.77)
4 DURBAN	***** *****	-.393490 (17.13)	-.443740 (13.44)	-.397440 (17.03)	-.392100 (17.02)	-.293670 (13.08)	-.369840 (15.99)
5 CURVE	***** *****	.26478 (15.34)	.25394 (10.45)	.26075 (15.00)	.242860 (14.20)	.254560 (15.20)	.26693 (15.37)
6 HERR	***** *****	.404720 (18.01)	.407270 (13.51)	.392950 (17.26)	.375640 (16.65)	.374080 (16.79)	.397090 (17.45)
7 STABLE	***** *****	.215000 (9.59)	.265380 (8.675)	.287710 (12.07)	.286430 (12.25)	.297370 (13.01)	.289080 (12.14)
8 YOUTH	***** *****	***** *****	***** *****	.066667 (4.02)	.018644 (1.146)	.026348 (1.652)	.085073 (5.15)
9 ALCAD	***** *****	***** *****	***** *****	.093045 (4.71)	.079714 (4.11)	.070307 (3.71)	.094859 (4.812)
10 BELT	***** *****	***** *****	***** *****	.094840 (4.63)	.084718 (4.20)	.109700 (5.56)	.101900 (4.983)
11 SURF	***** *****	***** *****	***** *****	-.170520 (8.32)	-.148740 (7.85)	-.142570 (7.68)	-.173010 (8.95)
CONSTANT	4.620500 (32.56)	6.598100 (29.36)	7.318400 (22.37)	6.664100 (29.47)	6.488000 (29.21)	-1.6256 (49.9)	4.9588 (33.41)
r-squared (Make/ Model Based)	.907	.9448	.9467	.9444	.5272	.6812	.9296
LIS (See Text)	1109	1832	*****	1907	1478	781	1856

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 (SF) is seen as the result of several observations:

1. Leaving SF and wheelbase (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).
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 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.

where is this shown?

When attention shifts 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 roll over is directly related to a "stability factor." The stability factor is defined as the ratio of one-half the track or tread width to the center of gravity height.

Based partly on the Robertson-Kelly study, Congressman T. Wirth (D-Co) petitioned 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 of 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 forty make/models and approximately 40,000 single vehicle accidents, including but not limited to fatals, from the CARDfile database. 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 assist them in enhancing 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.

2.0 TECHNICAL APPROACH

The approach that TSC used was to restructure the Harwin-Brewer (HB) CARDfile data on single vehicle accidents (SVA) so that a logit analysis could be performed at the accident level. The HB 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 4 million accidents from these states for 1983 through 1985 that were available for analysis (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 (Ref. 3) has indicated that CARDfile is representative of both national demographics and the accident experience. For a more detailed description of CARDfile, the reader is referred to the Harwin-Brewer study and to Reference 3.

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	3,982,117	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

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

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

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 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 HB 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 HB 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 (MSS) at NIH. The advantages of the MSS 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.

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		

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.

2.3 COMPUTER TECHNIQUE

The general approach to computer analysis of the data went through the following steps.

1. Bring the data-set onto the active hard disk from the MSS 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.

These three steps use SAS and presuppose computer system Job Control Language (JCL) in order to direct the system to the appropriate libraries and data-sets.

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. The r squared for each model was derived as explained in the methods section of this report.

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 non-vehicle 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

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

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 that of the one 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.

Besides 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 roll-overs 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:

$$L = C + A1*X1 + A2*X2 + \ldots + AN*XN \text{ where A1 is the coefficient of X1, etc.} \quad (1)$$

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 of 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 (graphical results in Figures 1-6 are also of importance). 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. Besides the variables in the first column are headings for two rows which give, respectively, the make/model based r^2 and the likelihood information statistic (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.

---

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