



ROBERT H. SMITH
SCHOOL OF BUSINESS

Leaders for the Digital Economy

Vehicle Accident Analysis

BUDT 733, Prof. Schmueli

Amelie Brandenburg

Jason Dell

Donna Donella

Rama Reddy





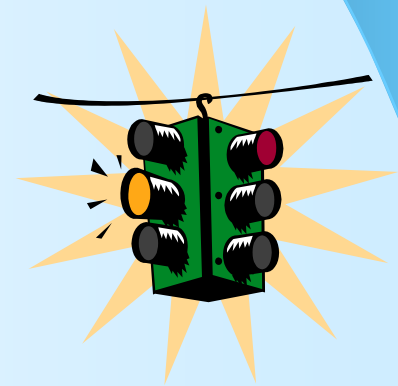
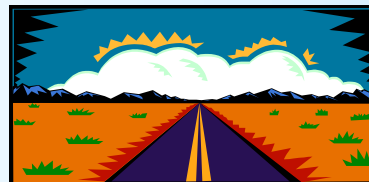
Outline

- Issue - Vehicular Accident Injuries
- Project Objective
- Data source and variables
- Research questions
- Methods of analysis
 - Exploratory analysis
 - Descriptive statistics
 - Logistic regression
 - Discriminant analysis
- Results
- Recommendations



Factors in Vehicular Accidents

- Physical environment
- Person – driver, passenger
- Vehicle related
- Other





Data Source

- The Bureau of Transportation Statistics (part of the U.S. Dept. of Transportation) gathers data on the estimated 6.4 million vehicle accidents reported each year. Data are accessible at <http://www.transtat.bts.gov>.
- The most recent sample of 55,000 records relates to 2001 vehicle accidents. From this data, we selected a random sample of 10,000 records for this project.



Objective

● Objective

– Evaluate the role of physical environment in vehicle accidents that result in injuries, specifically

- Profile variables of physical environment that increase the probability of injuries in vehicular accidents
- Predict the probability of injuries in accidents based on the selected variables of physical environment

● Hypothesis

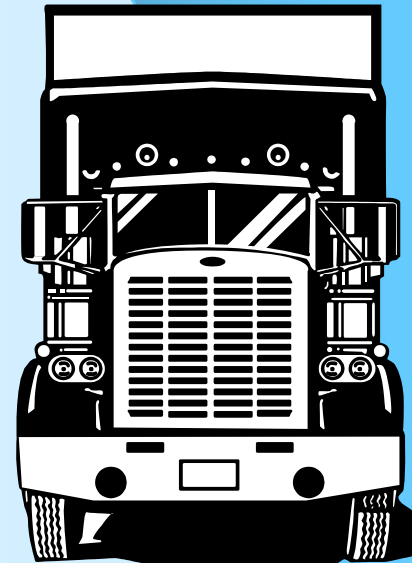
– Variables of physical environment have major influence on the probability of injury in vehicular accidents





Clients for this Study

- Federal Agencies and Dept. of Transportation
- State and County Governments
- Automobile Industry
- Insurance Industry – Health, Auto
- Businesses in road safety
- Commuters





Data - List of Variables

- Hour of accident - 0-24 hrs
- Alcohol
- Road alignment
- Manner of collision
- Traffic lanes - one /two/more
- Traffic conditions
- Road surface
- Population density
- Weekday
- Interstate Highway
- Relative to Roadway
- Work zone
- Relation to Junction
- Number of travel lanes
- Region - NE/MW/south/west
- Month
- Road surface conditions
- Traffic Flow
- Pedestrian/cyclist
- Light condition
- Speed limit
- Weather



Sample Data

CASENUM	STRATUM	WRK_ZONE	WKDY_I	EVENT1_I	NO_INJ_I	INJURY_CR ASH	MAXSEV_I	INT_HWY
110215646	1	0	7	25	3	1	3	0
110215716	3	0	6	25	1	1	2	0
110215725	1	0	3	25	3	1	1	0
110215728	2	0	2	25	0	0	0	0
110215741	1	0	1	25	0	0	0	0
110215744	4	0	3	25	0	0	0	0
110215764	1	0	1	25	4	1	2	0
110215767	2	0	6	45	0	0	0	0
110215782	1	0	5	25	0	0	0	0
110215791	1	0	7	37	3	1	1	0
110215808	4	0	1	45	1	1	1	0
110215823	1	0	1	37	1	1	1	0
110215824	4	0	7	25	1	1	1	0
110215880	1	0	3	25	0	0	0	0
110215887	1	0	7	37	0	0	0	0
110215889	4	0	7	45	0	0	0	0
110215896	4	0	3	25	0	0	0	0
110215925	1	0	3	35	1	1	1	0
110216025	1	0	5	38	0	0	0	0
110216026	4	0	1	25	0	0	0	0
110216061	1	0	2	25	4	1	1	0
110216072	4	0	7	24	0	0	0	0
110216073	1	0	3	38	0	0	0	0
110216134	1	0	4	25	0	0	0	0



Research Questions

Data are explored with focus on following questions

- What time of day the accidents result in most injuries?
- Are road alignment and profile factors in accident injuries?
- To what extent is speed limit a factor in injuries?
- Do seasonal factors – spring/summer/winter explain injuries?
- Whether manner of collision is a factor in the injury?
- Whether region affects the probability of injury?
- Do number of vehicles involved explain the injury?
- Do accidents on interstate highways result in more injuries



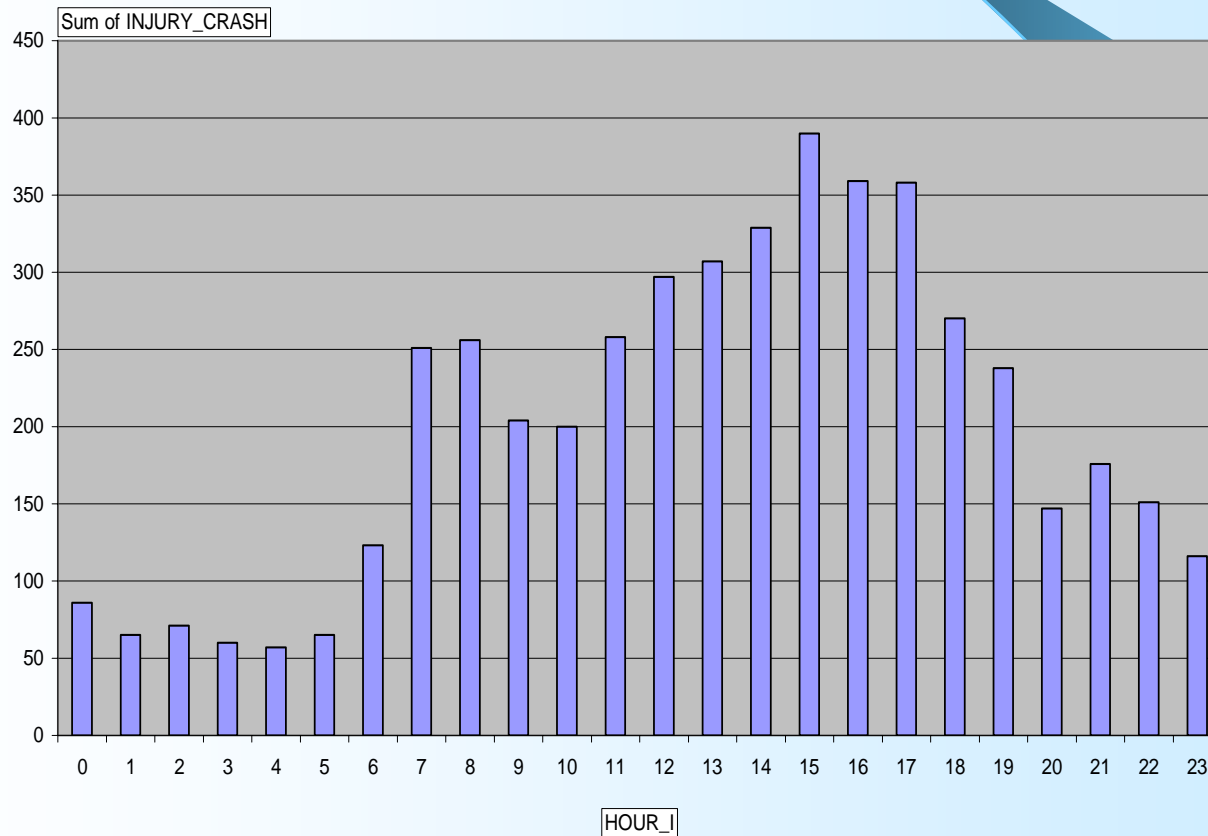
Methodology

- Started with data organization, exploratory analysis and descriptive statistics
- Selected variables for modeling the relationships.
- Logistic regression to predict the odds of injury
- Discriminant analysis to predict injury/non-injury group membership
- Evaluation of model results
- Recommendations from the analysis



Exploratory Analysis

Accidents involving injuries occur most frequently in the afternoon hours.





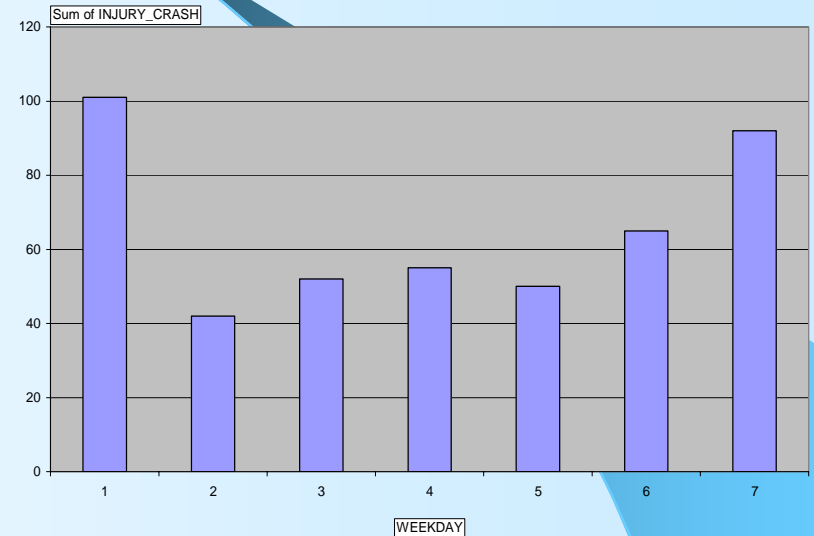
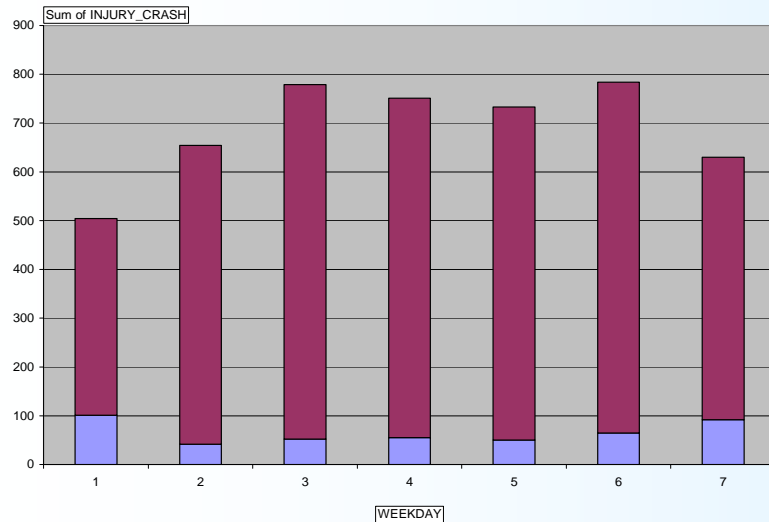
Injuries by Week and Month

MONTH	1	2	3	4	5	6	7	Total	Percent
1	41	47	72	72	46	46	38	362	7.5
2	31	37	59	65	58	54	63	367	7.6
3	32	51	61	57	81	69	67	418	8.6
4	52	56	52	55	50	73	49	387	8.0
5	41	51	82	83	67	59	43	426	8.8
6	38	61	61	46	50	80	50	386	8.0
7	48	72	89	48	70	52	32	411	8.5
8	60	50	57	74	76	74	51	442	9.1
9	43	45	64	57	56	64	63	392	8.1
10	38	82	59	88	55	71	52	445	9.2
11	45	46	63	45	59	84	58	400	8.3
12	35	56	60	61	65	58	64	399	8.3
Total	504	654	779	751	733	784	630	4835	100
Percent	10.4	13.5	16.1	15.5	15.2	16.2	13.0	100	



Exploratory Analysis

The day of the week doesn't seem to be a factor...

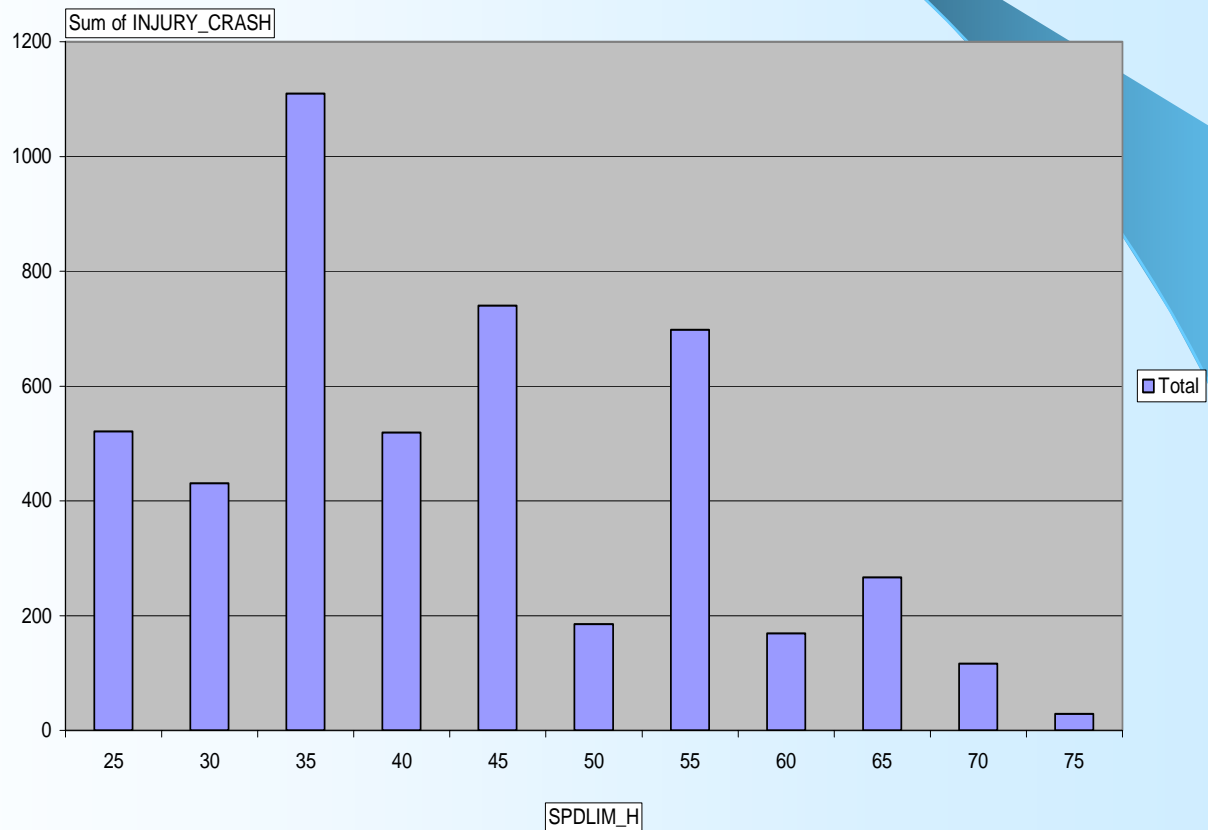


... but when you isolate alcohol related accidents a story begins to emerge.



Exploratory Analysis

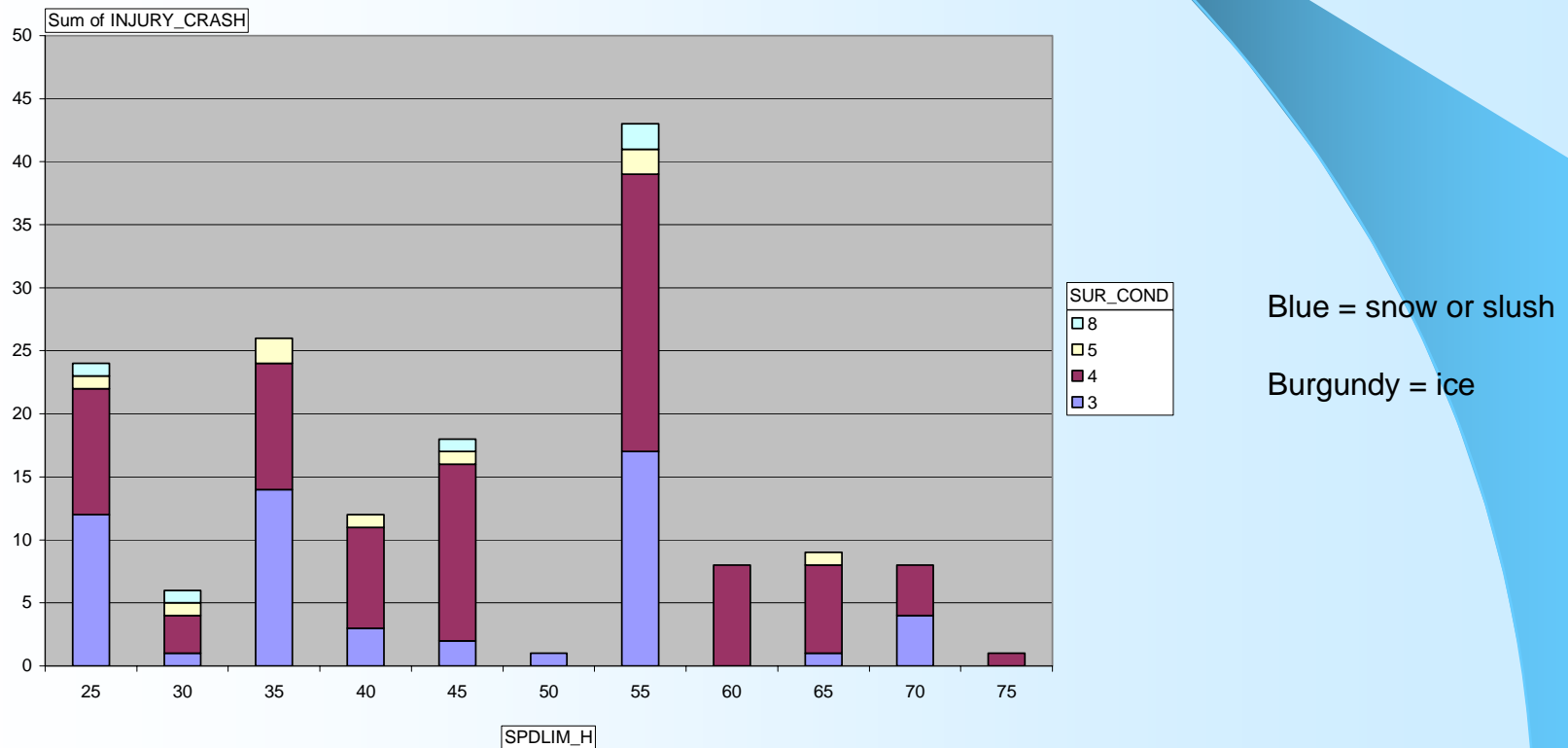
Most accidents occur at approximately 35 miles per hour.





Exploratory Analysis

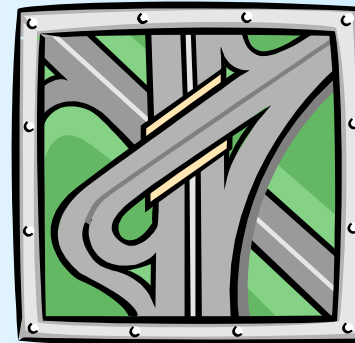
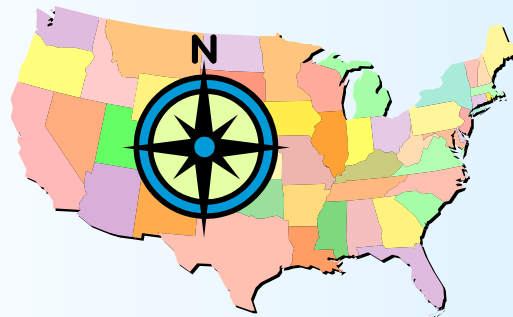
If you isolate accidents that occur on Slippery conditions, you see that most injuries occur at 55 MPH.





Variables chosen for modeling

- Hour – 6 hr intervals
- Road alignment
- Road profile
- Season
- Region
- Manner of collision
- Speed limit
- Vehicles involved



Modeling Strategy

- General to specific
 - Time, season, road, and collision
 - Time, season, region, road, and collision
 - Time, season and region interaction, road, and collision
- Model criteria
 - Reasonableness
 - Predictive accuracy
- Data preparation
 - Sample of 10,000 records
 - Initial variables – 28
 - Model variables – 12 to 15, dummy variables
 - Data partition – 60%/40%





Logistic Regression

- Qualitative response var. 1 = Injury 0 = No injury
- Independence of variables
- Profile of variables that best describes injury
- Prediction of injuries based on profiling
- Minimize misclassifications and maximize overall accuracy
- Cut off value 0.5





Logistic Regression

Model 1

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.63953114	0.165067	0.0001069	*
03_08	0.09918542	0.12917139	0.44257087	1.10427105
09_14	-0.0302764	0.12239034	0.80461746	0.97017735
15_20	0.11683763	0.05974631	0.05051673	1.12393689
ALIGN_I_2	0.14121759	0.07848415	0.07196909	1.15167522
Spring	0.21965012	0.08358181	0.00858972	1.24564087
Summer	0.24119872	0.08267291	0.00352843	1.27277398
Fall	0.11036064	0.07849389	0.15973081	1.11668074
PROFIL_I_3	0.45330581	0.17851323	0.01110618	1.57350528
SPDLIM_H	0.00188784	0.00206928	0.36160126	1.00188959
VEH_INVL	0.1346066	0.04126047	0.00110488	1.1440866
MANCOL_I_2-REAR	0.78871328	0.19452409	0.00005022	2.20056319
MANCOL_I_4-ANGLE	0.18385917	0.0603674	0.00232163	1.2018466

Training data

Error Report			
Class	# Cases	# Errors	% Error
1	2975	1558	52.37
0	3025	1209	39.97
Overall	6000	2767	46.12

Validation data

Error Report			
Class	# Cases	# Errors	% Error
1	1978	1055	53.34
0	2022	863	42.68
Overall	4000	1918	47.95



Logistic Regression Model 2

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.60903418	0.1725107	0.00041491	*
03_08	0.08632512	0.13006994	0.50689369	1.09016073
09_14	-0.05107668	0.12323972	0.67854476	0.9502058
15_20	0.11875071	0.06010022	0.04816858	1.12608922
ALIGN_I_2	0.09580129	0.07911576	0.22593363	1.1005404
Spring	0.23217715	0.0840885	0.00576062	1.26134312
Summer	0.24732943	0.08315539	0.00293654	1.28060091
Fall	0.1168991	0.07896433	0.13876548	1.12400603
REGION_2	-0.37093136	0.08665573	0.00001865	0.69009131
REGION_3	0.07461581	0.07952087	0.34808141	1.07747006
REGION_4	0.24985734	0.08981863	0.00540585	1.28384221
PROFIL_I_3	0.46062925	0.17959954	0.0103248	1.58507109
SPDLIM_H	0.00240776	0.00210802	0.25337532	1.00241065
VEH_INVL	0.11931346	0.04144797	0.00399404	1.12672305
MANCOL_I_2-REAR	0.80401832	0.19552338	0.0000392	2.23450184
MANCOL_I_4-ANGLE	0.19053856	0.06074547	0.00170878	1.20990098

Training data

Error Report			
Class	# Cases	# Errors	% Error
1	2975	1289	43.33
0	3025	1379	45.59
Overall	6000	2668	44.47

Validation data

Error Report			
Class	# Cases	# Errors	% Error
1	1978	887	44.84
0	2022	958	47.38
Overall	4000	1845	46.13



Logistic Regression

Model 3

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.50230372	0.15829232	0.00150734	*
03_08	0.06585259	0.13006741	0.61264902	1.06806922
09_14	-0.06159255	0.1232645	0.61730188	0.94026589
15_20	0.12273868	0.06008347	0.04107196	1.13058889
ALIGN_I_2	0.10584586	0.07905218	0.18059129	1.11165047
Spring*REGION_2	-0.24713884	0.11725279	0.03505315	0.78103226
Spring*REGION_3	0.26006782	0.09546112	0.00644316	1.29701805
Spring*REGION_4	0.4287546	0.13468824	0.00145601	1.53534424
Summer*REGION_2	0.07674035	0.11590154	0.50789642	1.07976162
Summer*REGION_3	0.11095104	0.09377426	0.23674114	1.11734021
Summer*REGION_4	0.20576884	0.12736782	0.10619207	1.22846925
Fall*REGION_2	-0.49264961	0.10451107	0.00000243	0.61100531
Fall*REGION_3	0.08282438	0.086313	0.33726576	1.08635104
Fall*REGION_4	0.40248448	0.11010651	0.00025677	1.49553573
PROFIL_I_3	0.46181905	0.17959623	0.01012796	1.58695817
SPDLIM_H	0.00222058	0.00209726	0.28968975	1.00222301
VEH_INVL	0.12032049	0.04149419	0.00373526	1.12785828
MANCOL_I_2-REAR	0.81687379	0.19556834	0.00002955	2.26341295
MANCOL_I_4-ANGLE	0.19331911	0.06072688	0.00145547	1.21326995

Training data

Error Report			
Class	# Cases	# Errors	% Error
1	2975	1353	45.48
0	3025	1294	42.78
Overall	6000	2647	44.12

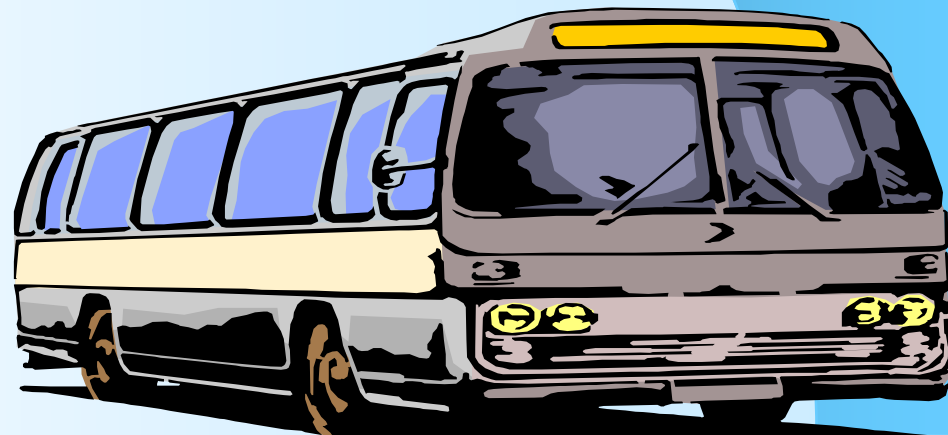
Validation data

Error Report			
Class	# Cases	# Errors	% Error
1	1978	922	46.61
0	2022	936	46.29
Overall	4000	1858	46.45



Discriminant Analysis

- Qualitative response var. 1 = Injury 0 = No injury
- Profile variables that best describes groups
- Given the variables predict membership to injury or non-injury groups
- Estimation of classification functions for two groups
- Minimize misclassifications and maximize overall accuracy
- Cut off value 0.5





Discriminant Analysis

Training data

Model 1

Variables	Classification Function	
	1	0
Constant	-20.6185379	-19.9792423
03_08	17.75182915	17.65247345
09_14	17.50257874	17.53284073
15_20	0.31071767	0.19373249
ALIGN_I_2	3.17435193	3.03376412
Spring	6.10405397	5.8843379
Summer	6.12920046	5.88793802
Fall	6.09499407	5.98459768
PROFIL_I_3	1.23662579	0.78991449
SPDLIM_H	0.26759824	0.26571327
VEH_INVL	2.99166417	2.85741496
MANCOL_I_2-REAR	1.73725379	0.97637361
MANCOL_I_4-ANGLE	1.30686736	1.12199306

Error Report			
Class	# Cases	# Errors	% Error
1	2975	1558	52.37
0	3025	1209	39.97
Overall	6000	2767	46.12

Error Report			
Class	# Cases	# Errors	% Error
1	1978	1055	53.34
0	2022	863	42.68
Overall	4000	1918	47.95

Validation data



Discriminant Analysis

Training data

Model 2

Variables	Classification Function	
	1	0
Constant	-22.255703	-21.646759
03_08	17.6318264	17.54524994
09_14	17.39513206	17.44648743
15_20	0.241785	0.12303382
ALIGN_I_2	3.3757205	3.28043652
Spring	6.06768608	5.83570671
Summer	6.14355516	5.89635324
Fall	6.06384802	5.94693995
REGION_2	5.20578146	5.57671213
REGION_3	4.68215132	4.60669661
REGION_4	5.56409597	5.312325
PROFIL_I_3	1.59064126	1.13673627
SPDLIM_H	0.24831736	0.24590679
VEH_INVL	2.94136548	2.82220197
MANCOL_I_2-REAR	1.95152771	1.17478144
MANCOL_I_4-ANGLE	1.36335731	1.17189658

Error Report			
Class	# Cases	# Errors	% Error
1	2975	1289	43.33
0	3025	1379	45.59
Overall	6000	2668	44.47

Error Report			
Class	# Cases	# Errors	% Error
1	1978	888	44.89
0	2022	959	47.43
Overall	4000	1847	46.18

Validation data



Discriminant Analysis

Training data

Model 3

Variables	Classification Function	
	1	0
Constant	-18.7803326	-18.2771034
03_08	17.65814972	17.59217262
09_14	17.49413681	17.55600166
15_20	0.30733269	0.18441325
ALIGN_I_2	3.17332721	3.06806183
Spring*REGION_2	2.7022512	2.95005918
Spring*REGION_3	2.32876444	2.0657599
Spring*REGION_4	2.96250701	2.53193378
Summer*REGION_2	2.77787805	2.70031691
Summer*REGION_3	2.10354948	1.99121737
Summer*REGION_4	2.70676374	2.49846292
Fall*REGION_2	2.83930182	3.32263947
Fall*REGION_3	2.12632442	2.04208088
Fall*REGION_4	3.08008218	2.67455697
PROFIL_I_3	1.33908427	0.88402832
SPDLIM_H	0.26123288	0.25900948
VEH_INVL	2.97946072	2.85942793
MANCOL_I_2-REAR	1.71974123	0.93042308
MANCOL_I_4-ANGLE	1.3390702	1.14463151

Error Report			
Class	# Cases	# Errors	% Error
1	2975	1352	45.45
0	3025	1294	42.78
Overall	6000	2646	44.10

Error Report			
Class	# Cases	# Errors	% Error
1	1978	921	46.56
0	2022	936	46.29
Overall	4000	1857	46.43

Validation data



Model Summary

- Results of Logistic and Discriminant analysis are consistent
- Error rate is high since we only focus on the variables of physical environment. Inclusion of person and vehicle variables may help in reducing the error rate
- Considering several qualitative variables the odds of logistic regression are more intuitive in explaining the relative significance of variables



Results Interpretation

- Physical environment variables predicts between 55% to 60% injuries that occur in vehicle accidents.
- You are more likely to meet with an accident that results in injury during evening rush hours.
- The odds of injuries in accidents are higher in spring and summer than those in fall and winter.
- Accidents on curved roads and on hilly terrain are more likely to result in injuries.
- You are less likely to meet with an accident that results in injury on Mondays



Results Interpretation

- The probability of injury is likely to be higher in accidents involving more than two vehicles.
- Accidents in the Western states are more likely to result in injuries.
- Rear and angular collisions are have greater probability of injury than those involving head on and side collisions.
- Speed limit within the context of physical environment variables is less significant.
- The accidents on interstate highways are less likely to results in injuries than those occurring on other roads.



Main Messages

- Injuries in accidents are likely to occur during evening rush hours.
- Injuries are likely in accidents that occur during spring and summer.
- You are more safer in the front passenger seat than in the back seat.
- Keep a safe distance from the front vehicle and pay attention to rear view mirror.
- Be careful when driving in California !



Recommendations

- Use large data and variables – cell phone use etc.
- Combine data on physical environment, person, and vehicle for better prediction
- Use these results in better road design and traffic management
- Periodically evaluate safety measures through *ex-post* analysis
- Collect more data on events leading to accidents



Questions ?

