

## Chapter 7: Categorical responses

7.1 For the degree of crash data, regard the event of interest as the occurrence of injury (fatal or non-fatal) in a crash. Develop a statistical model.

Firstly create a new variable `injured`, with `injured=0` when `degree=1` (non-casualty) and `injured=1` when `degree=2` (injury) or `degree=3` (fatal):

```
data injury;
set act.injury;
if degree=1 then injured=0;
else injured=1;
run;
```

The same variables as are significant for the ordinal regression, are significant here:

```
proc logistic data=injury descending;
class agecat sex roaduserclass/ param=ref;
model injured = roaduserclass agecat sex agecat*sex;
weight number;
run;
```

### The LOGISTIC Procedure

#### Model Information

Data Set	WORK.INJURY	
Response Variable	injured	
Number of Response Levels	2	
Weight Variable	Number	Number
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	209
Number of Observations Used	209
Sum of Weights Read	76341
Sum of Weights Used	76341

#### Response Profile

Ordered Value	injured	Total Frequency	Total Weight
1	1	132	32045.000
2	0	77	44296.000

Probability modeled is injured=1.

#### Class Level Information

Class	Value	Design Variables					
agecat	1	1	0	0	0	0	0
	2	0	1	0	0	0	0
	3	0	0	1	0	0	0
	5	0	0	0	1	0	0
	6	0	0	0	0	1	0
	7	0	0	0	0	0	1
	10	0	0	0	0	0	0

Sex	F	1		
	M	0		
roaduserclass	2	1	0	0
	4	0	1	0
	6	0	0	1
	10	0	0	0

## Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

## Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	103858.56	101068.62
SC	103861.91	101125.44
-2 Log L	103856.56	101034.62

## Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2821.9414	16	<.0001
Score	2609.8781	16	<.0001
Wald	1623.5450	16	<.0001

## Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
roaduserclass	3	1404.9916	<.0001
agecat	6	114.9136	<.0001
Sex	1	27.1318	<.0001
agecat*Sex	6	24.4544	0.0004

## Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.4683	0.0213	481.1289	<.0001
roaduserclass 2	1	0.1406	0.0270	27.0427	<.0001
roaduserclass 4	1	0.2558	0.0367	48.6248	<.0001
roaduserclass 6	1	2.8668	0.0776	1365.3552	<.0001
agecat 1	1	-0.1872	0.0326	33.0242	<.0001
agecat 2	1	-0.1169	0.0324	12.9869	0.0003
agecat 3	1	-0.0631	0.0368	2.9392	0.0865
agecat 5	1	0.0551	0.0303	3.3163	0.0686
agecat 6	1	0.0667	0.0334	3.9926	0.0457
agecat 7	1	0.1375	0.0347	15.7179	<.0001
Sex F	1	0.1751	0.0336	27.1318	<.0001
agecat*Sex 1 F	1	0.1353	0.0534	6.4174	0.0113
agecat*Sex 2 F	1	0.1236	0.0525	5.5423	0.0186
agecat*Sex 3 F	1	0.0509	0.0605	0.7065	0.4006
agecat*Sex 5 F	1	0.0283	0.0490	0.3338	0.5634
agecat*Sex 6 F	1	0.0162	0.0559	0.0844	0.7714
agecat*Sex 7 F	1	-0.1406	0.0598	5.5360	0.0186

## Odds Ratio Estimates

Point                      95% Wald

Effect	Estimate	Confidence Limits	
roaduserclass 2 vs 10	1.151	1.092	1.214
roaduserclass 4 vs 10	1.292	1.202	1.388
roaduserclass 6 vs 10	17.581	15.101	20.469

Association of Predicted Probabilities and Observed Responses

Percent Concordant	48.1	Somers' D	0.002
Percent Discordant	48.0	Gamma	0.002
Percent Tied	3.9	Tau-a	0.001
Pairs	10164	c	0.501

The striking result is that the odds of an injury for a motor cycle driver (road user class 6) is 17.6 times the odds of an injury for a car driver (road user class 10).

7.2 For the personal injury data, develop a statistical model for the occurrence of legal representation.

It only makes sense to include those variables that are known at the time of claim as predictors for legal representation. One may find, for example, that operational time is significant, but this would be as a result of the fact that claims being legally contested will take longer to settle.

In this data set, injury code is therefore the only available explanatory variable for legal representation. We could include accident year as an explanatory variable, in order to assess whether there is a trend in legal representation over time, but legal representation for the most recent claims will not have happened by the time the data was recorded. In practice, other aspects of the accident or claimant may be predictive. We examine legal representation by injury code:

```
proc freq data=act.persinj;
tables legrep*inj1 / norow nopercnt;
run;
```

Table of LEGREP by INJ1

LEGREP(LEGREP)	INJ1(INJ1)							
Frequency	1	2	3	4	5	6	9	Total
Col Pct								
0	5571 35.62	1152 34.12	374 33.01	56 29.63	85 45.21	121 47.27	649 51.67	8008
1	10067 64.38	2224 65.88	759 66.99	133 70.37	103 54.79	135 52.73	607 48.33	14028
Total	15638	3376	1133	189	188	256	1256	22036

Overall the percentage of legal representation is 64% (14028/22036). This varies from 48% (injury code 9) to 70% (injury code 4). As injury code 1 is by far the most frequent, it is declared as the base level:

```
proc genmod data=act.persinj descending;
class inj1 (ref="1") / param=ref;
model legrep = inj1 / dist=bin type3;
run;
```

## The GENMOD Procedure

## Model Information

Data Set	ACT.PERSINJ	
Distribution	Binomial	
Link Function	Logit	
Dependent Variable	LEGREP	LEGREP

Number of Observations Read	22036
Number of Observations Used	22036
Number of Events	14028
Number of Trials	22036

## Class Level Information

Class	Value	Design Variables					
INJ1	1	0	0	0	0	0	0
	2	1	0	0	0	0	0
	3	0	1	0	0	0	0
	4	0	0	1	0	0	0
	5	0	0	0	1	0	0
	6	0	0	0	0	1	0
	9	0	0	0	0	0	1

## Response Profile

Ordered Value	LEGREP	Total Frequency
1	1	14028
2	0	8008

PROC GENMOD is modeling the probability that LEGREP='1'.

## Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	22E3	28721.3092	1.3038
Scaled Deviance	22E3	28721.3092	1.3038
Pearson Chi-Square	22E3	22035.9999	1.0003
Scaled Pearson X2	22E3	22035.9999	1.0003
Log Likelihood		-14360.6546	

Algorithm converged.

## Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq	
Intercept	1	0.5917	0.0167	0.5590	0.6244	1255.56	<.0001	
INJ1	2	1	0.0661	-0.0122	0.1444	2.74	0.0980	
INJ1	3	1	0.1161	-0.0120	0.2441	3.15	0.0757	
INJ1	4	1	0.2733	-0.0406	0.5872	2.91	0.0879	
INJ1	5	1	-0.3996	0.1475	-0.6887	-0.1105	7.34	0.0067
INJ1	6	1	-0.4822	0.1263	-0.7297	-0.2347	14.58	0.0001
INJ1	9	1	-0.6586	0.0589	-0.7740	-0.5432	125.10	<.0001

```
Scale          0      1.0000      0.0000      1.0000      1.0000
```

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
INJ1	6	161.38	<.0001

Injury code is highly significant ( $p$ -value  $< 0.0001$ ).

7.3 The RTA also publish driver crash statistics by age, sex and blood alcohol concentration. This data set is provided on the companion website. Develop

(a) an ordinal regression model; and

(b) a nominal regression model

for degree of crash.

(a) **Ordinal regression model** Perform preparatory data manipulations and ordinal regression:

```
data bac;
set act.bac;
if number=0 or number=. then delete; /*take out zero frequencies*/
agecat=age;
if age=8 then agecat=7 ; /* Combine 60-69 and 70+*/
if age=4 then agecat=10; /* Make age 30-39 base level*/
if bac=1 then bac=10; /* Make bac=1 (legal) base level*/
run;

proc logistic data=bac;
class agecat sex bac / param=ref;
model degree = bac agecat sex agecat*sex ;
weight number;
run;
```

The interaction term `bac*agecat` is marginally significant ( $p$ -value = 0.0410) and has been omitted.

The LOGISTIC Procedure

Model Information

Data Set	WORK.BAC	
Response Variable	Degree	Degree
Number of Response Levels	3	
Weight Variable	Number	Number
Model	cumulative logit	
Optimization Technique	Fisher's scoring	
Number of Observations Read	174	
Number of Observations Used	174	
Sum of Weights Read	58890	
Sum of Weights Used	58890	

## Response Profile

Ordered Value	Degree	Total Frequency	Total Weight
1	1	63	34911.000
2	2	69	23351.000
3	3	42	628.000

Probabilities modeled are cumulated over the lower Ordered Values.

## Class Level Information

Class	Value	Design Variables					
agecat	1	1	0	0	0	0	0
	2	0	1	0	0	0	0
	3	0	0	1	0	0	0
	5	0	0	0	1	0	0
	6	0	0	0	0	1	0
	7	0	0	0	0	0	1
	10	0	0	0	0	0	0
Sex	f	1					
	m	0					
BAC	2	1	0	0	0		
	3	0	1	0	0		
	4	0	0	1	0		
	5	0	0	0	1		
	10	0	0	0	0		

## Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

## Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
92.2948	17	<.0001

## Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	85415.956	84952.633
SC	85422.274	85012.655
-2 Log L	85411.956	84914.633

## Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	497.3236	17	<.0001
Score	495.6705	17	<.0001
Wald	484.8243	17	<.0001

## Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
BAC	4	385.3662	<.0001
agecat	6	88.9992	<.0001
Sex	1	0.3955	0.5294

agecat\*Sex 6 23.9044 0.0005

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.3970	0.0230	297.5092	<.0001
Intercept	2	4.5721	0.0456	10071.2059	<.0001
BAC	2	-1.0029	0.2782	12.9949	0.0003
BAC	3	-0.9437	0.1381	46.7086	<.0001
BAC	4	-0.6432	0.0730	77.7037	<.0001
BAC	5	-1.2249	0.0764	256.9530	<.0001
agecat	1	0.2212	0.0356	38.5304	<.0001
agecat	2	0.1200	0.0360	11.1252	0.0009
agecat	3	0.0454	0.0410	1.2258	0.2682
agecat	5	-0.0344	0.0339	1.0288	0.3104
agecat	6	-0.0548	0.0375	2.1367	0.1438
agecat	7	-0.0806	0.0385	4.3840	0.0363
Sex	f	0.0240	0.0381	0.3955	0.5294
agecat*Sex 1 f	1	-0.2374	0.0599	15.7105	<.0001
agecat*Sex 2 f	1	-0.1955	0.0598	10.6708	0.0011
agecat*Sex 3 f	1	-0.0591	0.0694	0.7244	0.3947
agecat*Sex 5 f	1	-0.0948	0.0559	2.8748	0.0900
agecat*Sex 6 f	1	-0.1082	0.0637	2.8842	0.0895
agecat*Sex 7 f	1	0.00126	0.0671	0.0004	0.9851

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
BAC 2 vs 10	0.367	0.213 0.633
BAC 3 vs 10	0.389	0.297 0.510
BAC 4 vs 10	0.526	0.456 0.606
BAC 5 vs 10	0.294	0.253 0.341

Association of Predicted Probabilities and Observed Responses

Percent Concordant	45.2	Somers' D	-.064
Percent Discordant	51.6	Gamma	-.066
Percent Tied	3.2	Tau-a	-.042
Pairs	9891	c	0.468

As the test for proportional odds is rejected ( $p$ -value < 0.0001), the partial proportional odds model (Section 7.8) or the nominal regression model should be fitted.

### Partial proportional odds model

```

data bac2; set bac;
do i=1 to number;output;end;drop i;run;

data bac2; set bac2;
id=_n_;
do; if degree>2 then y=1; else y=0; degreej=2;output;end;
do; if degree>1 then y=1; else y=0; degreej=1;output;end;
run;

proc sort; by id; run;

proc genmod data=bac2;
class agecat bac sex degreej id;
model y = agecat sex bac agecat*sex degreej agecat*degreej sex*degreej
bac*degreej agecat*sex*degreej

```

```

      / dist=bin link=logit type3;
repeated subject=id / type=unstr;
run;

```

Examine the Type 3 tests:

Score Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
agecat	6	21.97	0.0012
Sex	1	23.25	<.0001
BAC	4	64.51	<.0001
agecat*Sex	6	5.17	0.5222
degreej	1	33.02	<.0001
agecat*degreej	6	9.73	0.1363
Sex*degreej	1	48.52	<.0001
BAC*degreej	4	17.09	0.0019
agecat*Sex*degreej	6	2.34	0.8857

As the interaction term `agecat*Sex*degreej` is not significant, remove it from the model and re-estimate:

```

proc genmod data=bac2;
class agecat bac sex degreej id;
model y = agecat sex bac agecat*sex degreej agecat*degreej sex*degreej bac*degreej
      / dist=bin link=logit type3;
repeated subject=id / type=unstr;
run;

```

Score Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
agecat	6	33.10	<.0001
Sex	1	20.92	<.0001
BAC	4	64.55	<.0001
agecat*Sex	6	23.39	0.0007
degreej	1	33.02	<.0001
agecat*degreej	6	14.93	0.0208
Sex*degreej	1	46.41	<.0001
BAC*degreej	4	17.18	0.0018

The proportional odds assumption is violated for age ( $p$ -value = 0.0208), sex ( $p$ -value < 0.0001) and BAC ( $p$ -value = 0.0018). This model is therefore retained. The table of parameter estimates is

Analysis Of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z	
Intercept	4.5137	0.0957	4.3261	4.7013	47.16	<.0001	
agecat	1	0.2093	0.1463	-0.0775	0.4961	1.43	0.1526
agecat	2	0.2680	0.1493	-0.0246	0.5606	1.80	0.0726
agecat	3	0.0843	0.1634	-0.2360	0.4046	0.52	0.6061
agecat	5	-0.0303	0.1327	-0.2903	0.2297	-0.23	0.8195
agecat	6	-0.2613	0.1401	-0.5359	0.0132	-1.87	0.0621
agecat	7	-0.4658	0.1386	-0.7375	-0.1941	-3.36	0.0008
agecat	10	0.0000	0.0000	0.0000	0.0000	.	.
Sex	f	0.6118	0.1024	0.4111	0.8125	5.97	<.0001
Sex	m	0.0000	0.0000	0.0000	0.0000	.	.
BAC	2	-0.7622	1.0227	-2.7666	1.2422	-0.75	0.4561



BAC	3		-1.5874	0.3279	-2.2300	-0.9447	-4.84	<.0001
BAC	4		-1.1280	0.2218	-1.5627	-0.6932	-5.08	<.0001
BAC	5		-1.8472	0.1608	-2.1623	-1.5321	-11.49	<.0001
BAC	10		0.0000	0.0000	0.0000	0.0000	.	.
agecat*Sex	1	f	-0.2366	0.0599	-0.3540	-0.1192	-3.95	<.0001
agecat*Sex	1	m	0.0000	0.0000	0.0000	0.0000	.	.
agecat*Sex	2	f	-0.1950	0.0598	-0.3123	-0.0777	-3.26	0.0011
agecat*Sex	2	m	0.0000	0.0000	0.0000	0.0000	.	.
agecat*Sex	3	f	-0.0594	0.0694	-0.1954	0.0765	-0.86	0.3916
agecat*Sex	3	m	0.0000	0.0000	0.0000	0.0000	.	.
agecat*Sex	5	f	-0.0963	0.0560	-0.2061	0.0134	-1.72	0.0853
agecat*Sex	5	m	0.0000	0.0000	0.0000	0.0000	.	.
agecat*Sex	6	f	-0.1115	0.0638	-0.2365	0.0135	-1.75	0.0804
agecat*Sex	6	m	0.0000	0.0000	0.0000	0.0000	.	.
agecat*Sex	7	f	-0.0038	0.0671	-0.1353	0.1276	-0.06	0.9546
agecat*Sex	7	m	0.0000	0.0000	0.0000	0.0000	.	.
agecat*Sex	10	f	0.0000	0.0000	0.0000	0.0000	.	.
agecat*Sex	10	m	0.0000	0.0000	0.0000	0.0000	.	.
degreej	1		-4.1151	0.0949	-4.3010	-3.9291	-43.37	<.0001
degreej	2		0.0000	0.0000	0.0000	0.0000	.	.
agecat*degreej	1	1	0.0108	0.1447	-0.2729	0.2944	0.07	0.9406
agecat*degreej	1	2	0.0000	0.0000	0.0000	0.0000	.	.
agecat*degreej	2	1	-0.1516	0.1478	-0.4414	0.1381	-1.03	0.3051
agecat*degreej	2	2	0.0000	0.0000	0.0000	0.0000	.	.
agecat*degreej	3	1	-0.0406	0.1613	-0.3568	0.2756	-0.25	0.8014
agecat*degreej	3	2	0.0000	0.0000	0.0000	0.0000	.	.
agecat*degreej	5	1	-0.0041	0.1313	-0.2615	0.2533	-0.03	0.9752
agecat*degreej	5	2	0.0000	0.0000	0.0000	0.0000	.	.
agecat*degreej	6	1	0.2113	0.1384	-0.0600	0.4826	1.53	0.1268
agecat*degreej	6	2	0.0000	0.0000	0.0000	0.0000	.	.
agecat*degreej	7	1	0.3952	0.1368	0.1270	0.6634	2.89	0.0039
agecat*degreej	7	2	0.0000	0.0000	0.0000	0.0000	.	.
agecat*degreej	10	1	0.0000	0.0000	0.0000	0.0000	.	.
agecat*degreej	10	2	0.0000	0.0000	0.0000	0.0000	.	.
Sex*degreej	f	1	-0.5967	0.0961	-0.7851	-0.4083	-6.21	<.0001
Sex*degreej	f	2	0.0000	0.0000	0.0000	0.0000	.	.
Sex*degreej	m	1	0.0000	0.0000	0.0000	0.0000	.	.
Sex*degreej	m	2	0.0000	0.0000	0.0000	0.0000	.	.
BAC*degreej	2	1	-0.2577	1.0296	-2.2756	1.7603	-0.25	0.8024
BAC*degreej	2	2	0.0000	0.0000	0.0000	0.0000	.	.
BAC*degreej	3	1	0.6994	0.3323	0.0481	1.3507	2.10	0.0353
BAC*degreej	3	2	0.0000	0.0000	0.0000	0.0000	.	.
BAC*degreej	4	1	0.5103	0.2227	0.0739	0.9467	2.29	0.0219
BAC*degreej	4	2	0.0000	0.0000	0.0000	0.0000	.	.
BAC*degreej	5	1	0.7083	0.1664	0.3822	1.0344	4.26	<.0001
BAC*degreej	5	2	0.0000	0.0000	0.0000	0.0000	.	.
BAC*degreej	10	1	0.0000	0.0000	0.0000	0.0000	.	.
BAC*degreej	10	2	0.0000	0.0000	0.0000	0.0000	.	.

Model equations are constructed as

$$\ln \frac{\hat{\pi}_j}{1 - \hat{\pi}_j} = \begin{cases} 0.3986 + 0.2201x_1 + \dots - 1.1389x_{17} & \text{for } j = 1 \text{ (non-casualty)} \\ 4.5137 + 0.2093x_1 + \dots - 1.8472x_{17} & \text{for } j = 2 \text{ (injury)} \end{cases}$$

The effect of BAC is shown in the table.  $\exp(\hat{\beta}_1)$  is the effect of the BAC level (compared to legal), on the odds of a crash being non-casualty; and  $\exp(\hat{\beta}_2)$  is the effect of the BAC level (compared to legal), on the odds of a crash being non-casualty or injury.

BAC	$\hat{\beta}_1$	$\hat{\beta}_2$	$\exp(\hat{\beta}_1)$	$\exp(\hat{\beta}_2)$
.020-.049	-0.7622	-1.0199	0.47	0.77
.050-.079	-1.5874	-0.888	0.20	2.01
.080-.149	-1.128	-0.6177	0.32	1.67
$\geq .150$	-1.8472	-1.1389	0.16	2.03

**(b) Nominal regression model**

```
proc logistic data=bac;
class agecat sex bac / param=ref;
model degree (ref=first) = bac agecat sex agecat*sex / link=glogit;
weight number;
run;
```

The same explanatory variables as for ordinal regression are significant.

## The LOGISTIC Procedure

## Model Information

Data Set	WORK.BAC	
Response Variable	Degree	Degree
Number of Response Levels	3	
Weight Variable	Number	Number
Model	generalized logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	174
Number of Observations Used	174
Sum of Weights Read	58890
Sum of Weights Used	58890

## Response Profile

Ordered Value	Degree	Total Frequency	Total Weight
1	1	63	34911.000
2	2	69	23351.000
3	3	42	628.000

Logits modeled use Degree=1 as the reference category.

## Class Level Information

Class	Value	Design Variables					
agecat	1	1	0	0	0	0	0
	2	0	1	0	0	0	0
	3	0	0	1	0	0	0
	5	0	0	0	1	0	0
	6	0	0	0	0	1	0
	7	0	0	0	0	0	1
	10	0	0	0	0	0	0
Sex	f	1					
	m	0					
BAC	2	1	0	0	0		
	3	0	1	0	0		
	4	0	0	1	0		
	5	0	0	0	1		
	10	0	0	0	0		

## Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

## Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
-----------	----------------	--------------------------

AIC	85415.956	84897.512
SC	85422.274	85011.238
-2 Log L	85411.956	84825.512

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	586.4445	34	<.0001
Score	686.1817	34	<.0001
Wald	601.4194	34	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
BAC	8	434.6839	<.0001
agecat	12	103.7009	<.0001
Sex	2	6.2950	0.0430
agecat*Sex	12	24.5469	0.0171

Analysis of Maximum Likelihood Estimates

Parameter	Degree	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	2	1	-0.4245	0.0233	332.3616	<.0001
Intercept	3	1	-4.0295	0.1054	1462.7506	<.0001
BAC 2	2	1	1.0145	0.2858	12.6004	0.0004
BAC 3	3	1	1.2716	1.0277	1.5311	0.2160
BAC 4	2	1	0.8369	0.1421	34.6840	<.0001
BAC 5	3	1	1.9881	0.3389	34.4142	<.0001
BAC 6	4	2	0.5892	0.0746	62.3733	<.0001
BAC 7	4	3	1.3942	0.2257	38.1527	<.0001
BAC 8	5	2	1.0735	0.0790	184.6167	<.0001
BAC 9	5	3	2.4065	0.1691	202.4469	<.0001
agecat 1	1	2	-0.2205	0.0360	37.4249	<.0001
agecat 2	1	3	-0.2789	0.1686	2.7370	0.0980
agecat 3	2	2	-0.1125	0.0364	9.5614	0.0020
agecat 4	2	3	-0.2995	0.1715	3.0524	0.0806
agecat 5	3	2	-0.0495	0.0415	1.4250	0.2326
agecat 6	3	3	-0.00579	0.1809	0.0010	0.9744
agecat 7	5	2	0.0341	0.0343	0.9856	0.3208
agecat 8	5	3	0.0440	0.1530	0.0829	0.7734
agecat 9	6	2	0.0414	0.0380	1.1862	0.2761
agecat 10	6	3	0.2912	0.1597	3.3248	0.0682
agecat 11	7	2	0.0532	0.0391	1.8565	0.1730
agecat 12	7	3	0.5194	0.1554	11.1748	0.0008
Sex f	f	2	-0.00531	0.0385	0.0190	0.8903
Sex m	f	3	-0.5297	0.2112	6.2922	0.0121
agecat*Sex 1 f	f	2	0.2370	0.0604	15.3813	<.0001
agecat*Sex 1 m	f	3	0.2596	0.3397	0.5840	0.4448
agecat*Sex 2 f	f	2	0.1921	0.0604	10.1274	0.0015
agecat*Sex 2 m	f	3	0.2269	0.3455	0.4312	0.5114
agecat*Sex 3 f	f	2	0.0748	0.0700	1.1445	0.2847
agecat*Sex 3 m	f	3	-0.4521	0.4388	1.0612	0.3029
agecat*Sex 4 f	f	2	0.0950	0.0565	2.8293	0.0926
agecat*Sex 4 m	f	3	0.1332	0.3036	0.1925	0.6608
agecat*Sex 5 f	f	2	0.1167	0.0644	3.2815	0.0701
agecat*Sex 5 m	f	3	0.0927	0.3282	0.0797	0.7777
agecat*Sex 6 f	f	2	0.0165	0.0679	0.0590	0.8080
agecat*Sex 6 m	f	3	-0.1309	0.3381	0.1500	0.6985

Odds Ratio Estimates

Effect	Degree	Point Estimate	95% Wald Confidence Limits
--------	--------	----------------	----------------------------

BAC	2	vs	10	2	2.758	1.575	4.829
BAC	2	vs	10	3	3.567	0.476	26.731
BAC	3	vs	10	2	2.309	1.748	3.051
BAC	3	vs	10	3	7.302	3.758	14.187
BAC	4	vs	10	2	1.802	1.557	2.086
BAC	4	vs	10	3	4.032	2.590	6.275
BAC	5	vs	10	2	2.926	2.506	3.416
BAC	5	vs	10	3	11.095	7.965	15.456

The effect of blood alcohol concentration is clear from the odds ratio estimates.

7.4 *In the Enterprise Miner data set, develop a statistical model for the occurrence of a claim.*

Variables in this data set are examined thoroughly in the document “SAS Miner preliminary analysis”.

After observing the diagrams and tables, we select the variables which look promising as explanatory variables for `claim_flag`. Fitting these explanatory variables in a binomial model with logit link, we then observe the Type 3 (LR) test results. If the variable is not significant in the type 3 test, it is eliminated from the model.

Note that we have observed that home value is associated with income. Even though those two variables are significant in type 3 test, we eliminate home value from the model. The final model is:

```
proc logistic data= exercise.claims_sas_miner descending;
class car_use car_type revoked married parent1 max_educ density / param=glm;
model claim_flag_2 = kidsdriv car_use bluebook retained car_type oldclaim
      clm_freq revoked mvr_pts income married parent1
      max_educ density
      / outroc=claims;
run;
```

The LOGISTIC Procedure

Model Information

Data Set	EXERCISE.CLAIMS_SAS_MINER
Response Variable	claim_flag_2
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	10303
Number of Observations Used	9733

Response Profile

Ordered Value	claim_flag_2	Total Frequency
1	1	2606
2	0	7127

Probability modeled is claim\_flag\_2=1.

NOTE: 570 observations were deleted due to missing values for the response or explanatory variables.

## Class Level Information

Class	Value	Design Variables					
CAR_USE	Commercial	1	0				
	Private	0	1				
CAR_TYPE	Panel Truck	1	0	0	0	0	0
	Pickup	0	1	0	0	0	0
	SUV	0	0	1	0	0	0
	Sedan	0	0	0	1	0	0
	Sports Car	0	0	0	0	1	0
	Van	0	0	0	0	0	1
REVOKED	No	1	0				
	Yes	0	1				
MARRIED	No	1	0				
	Yes	0	1				
PARENT1	No	1	0				
	Yes	0	1				
MAX_EDUC	<High School	1	0	0	0	0	
	Bachelors	0	1	0	0	0	
	High School	0	0	1	0	0	
	Masters	0	0	0	1	0	
	PhD	0	0	0	0	1	
DENSITY	Highly Rural	1	0	0	0		
	Highly Urban	0	1	0	0		
	Rural	0	0	1	0		
	Urban	0	0	0	1		

## Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

## Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	11311.882	8534.571
SC	11319.065	8706.969
-2 Log L	11309.882	8486.571

## Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2823.3108	23	<.0001
Score	2505.2386	23	<.0001
Wald	1808.7284	23	<.0001

## Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
KIDSDRIV	1	79.8707	<.0001
CAR_USE	1	151.3497	<.0001
BLUEBOOK	1	28.7073	<.0001
RETAINED	1	81.1130	<.0001
CAR_TYPE	5	132.7796	<.0001
OLDCLAIM	1	12.6512	0.0004
CLM_FREQ	1	63.4072	<.0001
REVOKED	1	103.1461	<.0001

MVR_PTS	1	68.0877	<.0001
INCOME	1	54.9708	<.0001
MARRIED	1	106.3467	<.0001
PARENT1	1	33.6430	<.0001
MAX_EDUC	4	52.5867	<.0001
DENSITY	3	860.1499	<.0001

## Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.1900	0.2259	0.7073	0.4003
KIDSDRIV	1	0.4688	0.0525	79.8707	<.0001
CAR_USE Commercial	1	0.8444	0.0686	151.3497	<.0001
CAR_USE Private	0	0	.	.	.
BLUEBOOK	1	-0.00002	4.333E-6	28.7073	<.0001
RETAINED	1	-0.0618	0.00686	81.1130	<.0001
CAR_TYPE Panel Truck	1	-0.0712	0.1322	0.2898	0.5904
CAR_TYPE Pickup	1	-0.0919	0.1125	0.6679	0.4138
CAR_TYPE SUV	1	0.0664	0.1116	0.3539	0.5519
CAR_TYPE Sedan	1	-0.6405	0.1126	32.3426	<.0001
CAR_TYPE Sports Car	1	0.3971	0.1264	9.8766	0.0017
CAR_TYPE Van	0	0	.	.	.
OLDCLAIM	1	-0.00001	3.622E-6	12.6512	0.0004
CLM_FREQ	1	0.2102	0.0264	63.4072	<.0001
REVOKED No	1	-0.8596	0.0846	103.1461	<.0001
REVOKED Yes	0	0	.	.	.
MVR_PTS	1	0.1036	0.0126	68.0877	<.0001
INCOME	1	-6.09E-6	8.213E-7	54.9708	<.0001
MARRIED No	1	0.6609	0.0641	106.3467	<.0001
MARRIED Yes	0	0	.	.	.
PARENT1 No	1	-0.5076	0.0875	33.6430	<.0001
PARENT1 Yes	0	0	.	.	.
MAX_EDUC <High School	1	0.4062	0.1405	8.3572	0.0038
MAX_EDUC Bachelors	1	-0.0575	0.1219	0.2229	0.6368
MAX_EDUC High School	1	0.3887	0.1288	9.1115	0.0025
MAX_EDUC Masters	1	-0.0778	0.1215	0.4097	0.5221
MAX_EDUC PhD	0	0	.	.	.
DENSITY Highly Rural	1	-1.8208	0.2002	82.6799	<.0001
DENSITY Highly Urban	1	1.1340	0.0574	389.8797	<.0001
DENSITY Rural	1	-1.6859	0.1227	188.9259	<.0001
DENSITY Urban	0	0	.	.	.

## Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
KIDSDRIV	1.598	1.442	1.771
CAR_USE Commercial vs Private	2.326	2.034	2.661
BLUEBOOK	1.000	1.000	1.000
RETAINED	0.940	0.928	0.953
CAR_TYPE Panel Truck vs Van	0.931	0.719	1.207
CAR_TYPE Pickup vs Van	0.912	0.732	1.137
CAR_TYPE SUV vs Van	1.069	0.859	1.330
CAR_TYPE Sedan vs Van	0.527	0.423	0.657
CAR_TYPE Sports Car vs Van	1.488	1.161	1.906
OLDCLAIM	1.000	1.000	1.000
CLM_FREQ	1.234	1.172	1.299
REVOKED No vs Yes	0.423	0.359	0.500
MVR_PTS	1.109	1.082	1.137
INCOME	1.000	1.000	1.000
MARRIED No vs Yes	1.936	1.708	2.196
PARENT1 No vs Yes	0.602	0.507	0.715
MAX_EDUC <High School vs PhD	1.501	1.140	1.977
MAX_EDUC Bachelors vs PhD	0.944	0.743	1.199
MAX_EDUC High School vs PhD	1.475	1.146	1.898
MAX_EDUC Masters vs PhD	0.925	0.729	1.174
DENSITY Highly Rural vs Urban	0.162	0.109	0.240
DENSITY Highly Urban vs Urban	3.108	2.777	3.478

DENSITY Rural vs Urban 0.185 0.146 0.236

Association of Predicted Probabilities and Observed Responses

Percent Concordant	82.6	Somers' D	0.654
Percent Discordant	17.2	Gamma	0.655
Percent Tied	0.2	Tau-a	0.256
Pairs	18572962	c	0.827

## Diagnostics

- **Area under curve**

The area under the curve is 0.827, which indicates good predictive ability. The ROC curve is shown in Figure 1.

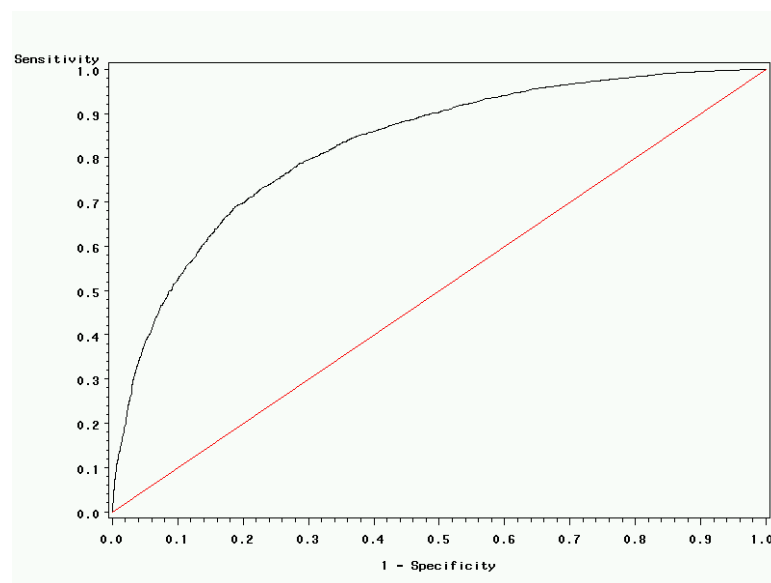


Figure 1: ROC curve, model for claim occurrence

- **Hat matrix diagonals**

The threshold value for the hat matrix diagonals is:  $h_{ii} > 2p/n = 2 * 24/9733 = 0.0049$ . 12% of the observations have  $h_{ii}$  higher than the threshold. It is worthwhile checking the validity of these cases.

**The model**

$$\ln \frac{\hat{\pi}}{1 - \hat{\pi}} = \exp\{0.5609 + 0.4705 * \text{kidsdriv} \\ + 0.0135 * \text{travtime} \\ - 0.000024 * \text{bluebook} \\ - 0.0619 * \text{retained} \\ - 0.2055 * \text{clm\_freq} \\ - 0.000012 * \text{oldclaim} \\ + 0.1025 * \text{mvr\_pts} \\ - 0.0000062 * \text{income} \\ + 0.8519 * (\text{car\_use} = \text{commercial}) \\ - 0.0787 * (\text{car\_type} = \text{panel truck}) \\ - 0.0999 * (\text{car\_type}=\text{pickup}) \\ + 0.0579 * (\text{car\_type}=\text{suv}) \\ - 0.6536 * (\text{car\_type}=\text{sedan}) \\ + 0.3847 * (\text{car\_type}=\text{sports car}) \\ + 0.6638 * (\text{not married}) \\ - 0.5309 * (\text{not single parent}) \\ + 0.3936 * (\text{max\_educ} < \text{high school}) \\ + 0.3746 * (\text{max\_educ}=\text{high school}) \\ - 0.0821 * (\text{max\_educ}=\text{bachelor}) \\ - 0.0972 * (\text{max\_educ}=\text{masters}) \\ - 2.0042 * (\text{density}=\text{highly rural}) \\ + 1.1035 * (\text{density}=\text{highly urban}) \\ - 1.7876 * (\text{density}=\text{rural})\}$$