536: Final Take-home Exam

Adrian Dobra
adobra@uw.edu

Problem 1 (50 points)

Table 1 refers to automobile accident records in Florida in 1988. It is a three-dimensional cross-classification of the following binary variables: "Safety Equipment in Use" with categories "Seat belt" and "None", "Whether Ejected" with categories "Yes" and "No" and "Injury" with categories "Non-fatal" and "fatal".

Safety Equipment in Use	Whether Ejected	Injury	
		Non-fatal	Fatal
Seat Belt	Yes	1105	14
	No	411111	483
None	Yes	4624	497
	No	157342	1008

Table 1: Automobile accident records in Florida.

Please analyze the data in Table 1 by answering the following questions:

Question 1. Consider all log-linear models associated with these three variables. Discuss their fit and choose a log-linear model that is representative for the associations among "Safety Equipment in Use", "Whether Ejected" and "Injury".

Question 2. Based on the log-linear model you selected at Question 1, derive the logistic regression:

$$\log \frac{P(\text{“Injury”} = \text{“Fatal”} | \text{“Safety Equipment in Use”}, \text{“Whether Ejected”})}{P(\text{“Injury”} = \text{“Non-fatal”} | \text{“Safety Equipment in Use”}, \text{“Whether Ejected”})}$$

Find the estimates of the coefficients of this logistic regression from the estimates of the u-terms of the log-linear model you selected at Question 1. What do you learn about your

data based on this regression?

Question 3. There are four logistic regression models that have "Injury" as the response variable and "Safety Equipment in Use", "Whether Ejected" as possible explanatory variables. Fit these models directly (using the function "glm") and discuss which of the four models you consider most appropriate for these data. Please note that you do not necessarily need to settle on the regression model you found at Question 2.

Question 4. Summarize your findings. What seem to be the factors determining the seriousness of the injuries sustained after a car accident? What are is the relationship between these factors? Which seems to be the most relevant factor?

Problem 2 (50 points)

Table 2 contains a 2^6 contingency table that cross-classifies binary risk factors denoted by A, B, C, D, E, F for coronary trombosis from a prospective epidemiological study of 1841 workers in a Czechoslovakian car factory. Here A indicates whether or not the worker "smokes", B corresponds to "strenuous mental work", C corresponds to "strenuous physical work", D corresponds to "systolic blood pressure", E corresponds to "ratio of β and α lipoproteins", and F represents "family anamnesis of coronary heart disease". Identify relevant loglinear models for the data in Table 2. Discuss your rationale for choosing your preferred models. Based on your top loglinear model, which of the variables A, B, C, D, E are directly determinant of variable F? In other words, which of the variables A, B, C, D, E would you choose to include in a logistic regression model for variable F, and which variables would you choose to drop from this regression?

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
neg	< 3	< 140	no		44	40	112	67
			yes		129	145	12	23
		≥ 140	no		35	12	80	33
			yes		109	67	7	9
	≥ 3	< 140	no		23	32	70	66
			yes		50	80	7	13
		≥ 140	no		24	25	73	57
			yes		51	63	7	16
pos	< 3	< 140	no		5	7	21	9
			yes		9	17	1	4
		≥ 140	no		4	3	11	8
			yes		14	17	5	2
	≥ 3	< 140	no		7	3	14	14
			yes		9	16	2	3
		≥ 140	no		4	0	13	11
			yes		5	14	4	4

Table 2: Prognostic factors for coronary heart disease as measured on Czech autoworkers.