Additional Exercise 4.3

Data: presence or absence of disease (liver cirrhosis) of patients in two cities and classified as alcoholics/non-alcoholics. The most natural notation is,

$$\begin{aligned} y_{ijk} &= \text{presence (1) or absence (0) of disease for } k\text{'th person in city } i \text{ and in alcoholic group } j, \\ &\quad i = 1, 0 \text{ (New York, Philadelphia)},\ j = 1, 0 \text{ (alcoholic, non-alcoholic)},\ k = 1, ..., n_{ij}. \end{aligned}$$

An alternative notation would just use y_i, with i referring to patient number, $i = 1, \ldots, 4600$.

Statistical models:
The data are grouped with four groups, corresponding to the combinations of city and alcoholic status, and the full model for grouped data simply assumes independent binomial distributions for the four groups. We could phrase this as,

$$\Pr(y_{ijk} = 1) = p_{ij}, \text{ with no restrictions on the } p_{ij}\text{'s.}$$

The additive logistic regression model imposes the restriction that

$$\text{logit}(p_{ij}) = \mu + \alpha_i + \beta_j.$$

Effectively, this model assumes no interaction on logistic scale between effects of city and alcoholic groups. Adding the interaction takes us back to the full model without any reduction in the description of the data. The Minitab output from fitting those two models is shown on the next page.

Interpretation of results:
The deviance and Pearson goodness-of-fit statistics for the additive model are both non-significant, indicating no evidence of interaction between effects of city and alcohol group. The Hosmer-Lemeshow test is less useful here (its main use is for ungrouped data). We also see that the interaction term in the full model is non-significant (with the same P-value), this is an equivalent way of assessing model fit for the additive model.

In the additive model, the z-tests for city and alcohol are both clearly significant, so we do not bother calculating the likelihood ratio statistics (by fitting the relevant submodels and computing the differences in deviance).

The odds-ratio for city (Philadelphia vs. New York) is 0.51, which means that the risk (odds) of being diseased is about twice as high in the patient group in New York than in Philadelphia. It is not clear what that really means, because it may very well relate to the selection of the patients. Maybe cities should be considered as blocks, and not be given any particular interpretation.

The odds-ratio for alcohol group (alcoholic vs. non-alcoholic) is 9.1, which means that the risk (odds) for disease is much higher in the alcoholic group. Maybe not surprising, with today's understanding, but these data are from before 1942.

HS04_3GRP.CSV

Binary Logistic Regression: diseased versus city, alcoholic

Method

Link function	Logit
Categorical predictor coding	(1, 0)
Rows used	4

Response Information

Variable	Value	Count	Event Name
diseased	Event	210	sick
	Non-event	4390	
total	Total	4600	

Regression Equation

P(sick) = exp(Y')/(1 + exp(Y'))

Y' = -2.886 + 0.0 city_New York - 0.681 city_Philadelphia + 0.0 alcoholic_0
+ 2.203 alcoholic_1

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-2.886	0.164	-17.61	0.000	
city					
Philadelphia	-0.681	0.172	-3.97	0.000	1.04
alcoholic					
1	2.203	0.160	13.73	0.000	1.04

Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
city			
Philadelphia	New York	0.5059	(0.3614, 0.7082)
alcoholic			
1	0	9.0565	(6.6127, 12.4035)

Odds ratio for level A relative to level B

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	1	0.25	0.618
Pearson	1	0.25	0.618
Hosmer-Lemeshow	1	0.10	0.750

Analysis of Variance

		Wald Test	
Source	DF	Chi-Square	P-Value
Regression	2	233.39	0.000
city	1	15.77	0.000
alcoholic	1	188.58	0.000

HS04_3GRP.CSV

Binary Logistic Regression: diseased versus city, alcoholic

Method

Link function	Logit
Categorical predictor coding	(1, 0)
Rows used	4

Response Information

Variable	Value	Count	Event Name
diseased	Event	210	sick
	Non-event	4390	
total	Total	4600	

Regression Equation

P(sick) = exp(Y')/(1 + exp(Y'))

Y' = -2.944 + 0.0 city_New York - 0.609 city_Philadelphia + 0.0 alcoholic_0
+ 2.325 alcoholic_1 + 0.0 city*alcoholic_New York 0 + 0.0 city*alcoholic_New York 1
+ 0.0 city*alcoholic_Philadelphia 0 - 0.175 city*alcoholic_Philadelphia 1

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-2.944	0.205	-14.35	0.000	
city					
Philadelphia	-0.609	0.228	-2.67	0.008	1.81
alcoholic					
1	2.325	0.293	7.93	0.000	3.45
city*alcoholic					
Philadelphia 1	-0.175	0.351	-0.50	0.618	3.59

Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
city			
Any level	Any level	*	(*, *)
alcoholic			
Any level	Any level	*	(*, *)

Odds ratio for level A relative to level B
Odds ratios are not calculated for predictors that are included in interaction terms because these ratios depend on values of the other predictors in the interaction terms.

Analysis of Variance

		Wald Test	
Source	DF	Chi-Square	P-Value
Regression	3	235.00	0.000
city	1	7.14	0.008
alcoholic	1	62.83	0.000
city*alcoholic	1	0.25	0.618