Topics to be covered this week

STA 138

Monday, Oct 22 Three-way tables (chap 2.7 in the text, and Handout 7)

Wednesday, Oct 24 Logistic and Poisson Regression models (chaps 3.1-3.3 and Handout 9)

Friday, Oct 26 Rate Regression, Negative binomial regression, Generalized Linear models (chap 3.3 in the text and Handout 10).

Homework 4: (Due on Monday, October 29)

You may form a group of 3 students registered in this course and submit one completed homework for the group. The front page should display only the names of the students in the group. The actual work should start from the second page.

1. The following is data on the death penalty in Florida from 1981.

		Death Penalty	
Victims Race	Defendants Race	Yes	No
White	White	19	132
	Black	11	52
Black	White	1	9
	Black	6	97

(a) For each of victims race, calculate the risk (probability) of death penalty given each of defendants race.

(b) Calculate and display the marginal table (2×2 table) combining over victims race.

(c) Using the marginal table, calculate the risk of death penalty given each of defendants race.

(d) Is this an example of Simpson's paradox? Explain.

2. In a study on CHD (coronary heart disease), 200 individuals from each of the six age groups were taken, and presence or absence of CHD of each sample individual was recorded. The following data were obtained.

	CHD	
Age	Present	Absent
[25,35]	10	190
[35,45]	23	177
[45,55]	71	129
[55,65]	133	67
[65,75]	179	21
[75,85]	191	6

The model is $\pi'(X) = \beta_0 + \beta_1 X$, where X = age, $\pi(X)$ is the proportion with CHD present at age X, and $\pi'(X)$ is the logit of $\pi(X)$.

A logistic regression model was fitted using the R command:

glm(cbind(Present,Absent)~Age,family='binomial'),

and the following were obtained:

$\hat{\beta}_0 = -7.0.593$, $\hat{\beta}_1 = 0.12918, s(\hat{\beta}_0) = 0.38350, s(\hat{\beta}_1) = 0.00685$.

(a) Obtain the sample proportion of individuals who have CHD for each of the age groups.

(b) Plot the sample proportions as well as the fitted proportions against age X on the same graph. Also plot the logit of sample proportions as well as fitted logistic regression against age on the same graph. Summarize your findings.

(c) Obtain an approximate 99% confidence interval for β_1. Use this confidence interval to test the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at level $\alpha = 0.01$.

(d) It is desired to obtain an approximate confidence interval for $\pi(67)$. The SE for logit($\hat{\pi}(67)$) turns out to be 0.11757. Use this information to obtain an approximate 99% confidence interval for $\pi(67)$. [Hint: First obtain a 99% confidence interval for logit($\pi(67)$).]

3. A researcher was interested in determining risk factors for high blood pressure (hypertension) among women. Data from a random sample of 680 women were collected to find out if they were suffering from hypertension. Each sample individual's smoking habit and age were also recorded. A summary of the data is given below (without the age)

	Hypertension	
	Yes	No
Smokers	28	271
Nonsmoker	13	368
Total	41	639

Let π be the probability of having hypertension. and π' be the logit of π. The researcher decided to fit a logistic regression model

$$\pi' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$

where X_1=smoking status (1=smoker,0=nonsmoker), and X_2=age, The following summary was obtained

Pararameter	Estimate	SE
β_0	-2.8	1.2
β_1	0.706	0.311
β_2	0.0004	0.0001
β_3	0.0006	0.0004

(a) What is the estimated logistic regression model for the relationship between age and hypertension for nonsmokers? Repeat this for smokers.

(b) Estimate a 25-year old smoker's probability of having hypertension.

(c) Estimate the odds ratio for hypertension comparing a 25-year old smoker to a 26 year old smoker. Interpret the ratio.

(d) Since $z = \hat{\beta}_3/s(\hat{\beta}_3) = 1.5$ is rather small in magnitude, we can drop the interaction term from the logistic model (when testing $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$). So we may consider a logistic regression model which has variables X_1 and X_2, but no interaction term X_1X_2. For this model, show that the odds ratio for hypertension comparing a 25-year old smoker to a 26 year old smoker does not depend on the smoking status. Find an expression for this odds ratio in terms of the parameters of the model. [You cannot find a numerical value for this odds ratio since we have not fitted the model without the interaction term.]

4. An experiment analyzes imperfection rates for two processes used to fabricate silicon wafers for computer chips. For treatment A applied to 10 wafers, the number of imperfections are 8,7,6,6,3,4,7,2,3,4. Treatment B applied to 10 other wafers has 9,9,8,14,8,13,11,5,7,6 imperfections. Treat the counts as independent Poisson variate having means μ_A and μ_B. Consider the model $\log \mu = \beta_0 + \beta_1 X$, where $X = 1$ for treatment B and $X = 0$ for treatment B.

A part of the computer output is given

Parameter	Estimate	SE
β_0	1.6094	0.1414
β_1	0.5878	0.1764

(a) Show that $\beta_1 = \log \mu_B - \log \mu_A$ and $e^{\beta_1} = \mu_B/\mu_A$.

(b) Write down the fitted model and interpret $\hat{\beta}_1$.

(c) Carry out a test for $H_0 : \mu_A = \mu_B$ against $H_1 : \mu_A \neq \mu_A$ at a level $\alpha = 0.05$.

(d) Construct an approximate 95% confidence interval for μ_B/μ_A.