BIOSTAT 651
Homework #3
due: Monday, March 13

- turn in at the start of class

- each sub-question=2 points; total 20 points

1. A retrospective cohort study was carried out to study factors affecting chronic obstructive pulmonary disease (COPD) risk. The observed data consist of triplets (Y_i, S_i, P_i), where Y_i=1 for subjects with COPD (0 otherwise); S_i=1 for smokers (0 for non-smokers) and P_i is an indicator for residence in a zip code considered to be highly polluted. The total sample size was $n = 200$, with the observed data summarized by the following tables:

   ○ for non-smokers ($S_i = 0$):

	Y_i=0	Y_i=1	total
P_i=0	30	20	50
P_i=1	40	10	50
total	70	30	100

   ○ for smokers ($S_i = 1$):

	Y_i=0	Y_i=1	total
P_i=0	20	30	50
P_i=1	18	32	50
total	38	62	100

(a) Fit the following model:

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_0 + \beta_1 S_i + \beta_2 P_i + \beta_3 S_i P_i.$$

and estimate β_0, β_1, β_2 and β_3, where $\pi_i = P(Y_i = 1|S_i, P_i)$. It should be done by hand, not in computer.

(b) Provide interpretations for $\exp\{\widehat{\beta}_0\}$, $\exp\{\widehat{\beta}_2\}$ and $\exp\{\widehat{\beta}_3\}$.

(c) Carry out the likelihood ratio test for $\beta_2 = \beta_3 = 0$ using Deviance. Please write down full and reduced models and their deviances, LRT test statistics and conclusion. You can use statistical software (e.g SAS or R) to solve this problem.

2. Solve 7.2 (a) and (b) on page 144 of the textbook (Dobson).

3. A retrospective study was carried out by the University of Adelaide on a random sample of graduate students. Each student was followed for 50 years after graduation and classified as dead or alive. Data are contained in the file *Adelaide.txt*, with columns YEAR, DEPT, SURVIVORS, TOTAL. The following model is of interest:

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_0 + \beta_1(YEAR_i - 1900) + \beta_2 ART_i + \beta_3 MED_i + \beta_4 ENG_i,$$

where $YEAR_i$ is the year of graduation; $ART_i = 1$ if the student graduates from the Arts Department and 0 otherwise; MED_i and ENG_i are defined analogously. Note that $\pi_i = \pi(\mathbf{x}_i) = P(Y_i = 1|\mathbf{x}_i)$ with $Y_i = 1$ if the graduate survives (0 if not). You can use SAS (or R) to solve this problem.

(a) Estimate and Interpret β_0, β_1 and β_2.

(b) [4 pt] Carry out model diagnostics for this analysis. Specifically

1) draw residual, leverage and cook's distance plots and determine whether there are any outliers.

2) calculate a variance inflation factor (VIF) for each independent variable and determine whether there is a multicollinearity problem.

3) calculate the pseudo R^2.

4) carry out the Hosmer-Lemeshow goodness of fit test.

Does this model fit the data well? please justify your answer.

4. Using the table below, determine log-likelihood and score functions for the model

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \alpha + \beta x,$$

where $\pi_i = P(Y = 1|X)$.

	Y=0	Y=1	Total
X=0	n_{00}	n_{01}	n_0
X=1	n_{10}	n_{11}	n_1

(a) Obtain MLE of β, $\widehat{\beta}$, and show that $\exp(\widehat{\beta})$ is the same as the sample odds ratio, $n_{00}n_{11}/(n_{01}n_{10})$.

(b) Show that the asymptotic variance estimate of log OR (i.e. $\widehat{\beta}$) is

$$\widehat{Var}(\widehat{\beta}) = \frac{1}{n_{00}} + \frac{1}{n_{11}} + \frac{1}{n_{01}} + \frac{1}{n_{10}}$$