# STAT 425/525 - Homework 8
# Due Monday November 7, 2016

*Unless otherwise indicated you can use computer software to do the following problems.*

1. Diabetes melitus is a major public health problem. Patients who have had the disease longer than 20 years tend to develop diabetic retinopathy (damage to blood vessels in the retina). Early detection is crucial to minimizing damage. A study was conducted to determine if general practitioners could be trained to detect diabetic retinopathy after a short training session. The eye skills of 85 general practioners were evaluated before and after training workshops. Each doctor was asked to evaluate 4 patients who did and 4 patients who did not have retinopathy before and after the workshop. We will look at the results for those who did not have retinopathy, i.e. we will look at the evidence for an improvement in specificity. Let $X$ denote post-workshop with $X = 1$ denoting not satisfactory and $X = 2$ denoting satisfactory. Let $Y$ denote pre-workshop with $Y = 1$ denoting not satisfactory and $Y = 2$ denoting satisfactory.

| | $X = 1$ | $X = 2$ | |
|---|---|---|---|
| $Y = 1$ | 15 | 50 | 65 |
| $Y = 2$ | 5 | 15 | 20 |
| Total | 20 | 650 | 85 |

   (a) Let $p_1$ be the proportion of non-satisfactory results pre-workshop and and $p_2$ be the proportion of non-satisfactory results post-workshop. Estimate the difference in proportions $p_1 - p_2$ and give a standard error for that estimate.

   (b) Construct an approximate 95% confidence interval for $p_1 - p_2$ and interpret it.

   (c) Test the hypothesis that the proportions of non-satisfactory results is equal pre and post-workshop.

2. To study an association between the risk of a low birth weight newborn and maternal smoking, an infant who weighed less than 2500 grams at birth (case) was matched to an infant whose birth weight was greater than 2500 grams at birth (control) so that the mother of each infant had the same prepregnancy weight. The risk factor is the mother's smoking exposure ($E =$ smoker, $\overline{E}$ =nonsmoker). A total of $n = 167$ matched pairs (matched on prepregnancy weight) was included i the study. The table below contains the results

| | $< 2500\ E$ | $< 2500\ \overline{E}$ | |
|---|---|---|---|
| $\geq 2500\ E$ | 15 | 22 | 37 |
| $\geq 2500\ \overline{E}$ | 40 | 90 | 130 |
| Total | 55 | 112 | 167 |

Fit a logistic regression model to these data. Give an estimate of the ratio of the odds that an infant exposed to smoking is low birthweigh to the odds that an infant not exposed is low birthweight. Give an approximate 95% confidence interval and interpret the interval. R-code for this analysis will be provided in a separate script file.

3. A data set lipcancer.txt is attached. The data set contains the observed numbers of lip cancer cases (obs) in 56 Scottish districts between 1975-1980, the expected number of cases (exp), the percentage of the district population employed in agriculture, fishing, and forestry (aff), and the latitude (lat) and longitude (long) coordinate of the center of each district. The observed counts by themselves can be misleading. For example, district 33 had 7 observed cases but the expected number of cases was also 7. But district 8 had 7 observed cases with only 2.3 cases expected. The ratio of observed to expected counts is a Standardized Morbidity Ratio (SMR). We will model the SMR using a Poisson rate model with the number of expected cases included as an offset term in the model.

   (a) Fit a Poisson regression model with observed number of cases as the response and aff as the explanatory variable. Summarize the results. Show me the output. Interpret the estimated coefficient associated with aff. Give an approximate 95% confidence interval for that parameter.

   (b) One possible confounding issue is how far north a district lies. More northern districts are more rural and can be expected to have a higher percentage of people who work outside. But the farther north one lives the less the sun exposure. Fit a Poisson regression model with observed number of cases as the response and aff and lat as explanatory variables. lat accounts for the northingness of the district. Summarize the results. Show me the output. Interpret the estimated coefficient associated with aff after accounting for the possible confounding effects of lat. Give an approximate 95% confidence interval for that parameter.

4. Refer to the $(AC, AM, CM)$ model from the Alcohol-Cigarette-Marijuana use example. On page 211 in the notes I gave the estimated conditional odds ratio of alcohol use among cigarette smokers to alcohol use among non-smokers $\widehat{\theta}_{AC|M} = 7.8$. I also gave approximate 95% (Wald) confidence intervals for $\theta_{AM|C}$ and $\theta_{CM|A}$. Give the points estimate for these quantities for these latter 2 odds ratios and confirm the intervals.