Applied Logistic Regression
Exercises 4

1. Consider the data on sentiments towards housing integration in Minnesota (used in Homework 3). Fit a logistic regression model to the data that has the main effects for contacts and norms. Compute both the Pearson residuals and the deviance residuals for the model. Display the results in a single matrix that has the observed counts, expected counts under the estimated model, and the residuals. Do the results suggest lack of fit?

2. Carry out a simulation experiment to study the effect of variable centering on the sampling distribution of the coefficient estimates. The setting is as follows: $Y_i \sim Ber(p_i), i = 1, \ldots, n$, where $logit(p_i) = \beta_0 + \beta_1 G_i + \beta_2 x_i$, where $n = 50$; $G_i = 1$ for $i = 1, \ldots, 20$ and $G_i = 0$ for $i = 21, \ldots, 50$; and $x_i = i/n$. Repeat the following $N = 300$ times: (i) generate values for the Bernoulli variables; (ii) compute the MLEs of the regression coefficients with the glm-function; (iii) store the coefficients into an $N \times 3$ matrix. Switching from the original x-values to the centered ones will not change $\hat{\beta}_2$ but the intercept changes to $\hat{\beta}_0 + \hat{\beta}_2 \times 25.5/n$. Use $\beta_0 = -0.9, \beta_1 = 0.3, \beta_2 = 1$ as true values in simulation. As an answer to this assignment, give the R code you used; give scatter plots of $(\hat{\beta}_0, \hat{\beta}_2)$ in both cases. What practical implication does the result have in applications?

3. The excel-file lowbtm11.xls contains data on 56 matched case-control pairs. The interest centers on what has caused a baby to have a low birth weight $<$ 2500 g (LOW = 1) or not (LOW = 0). The actual weight (in pounds) in given in LWT, but is not to be used in this exercise. Also, AGE has been used as a matching variable, so it cannot be used as an explanatory variable. Use conditional logistic regression to study the effect of the other variables: RACE (1 = white; 2 = black; 3 = other), SMOKE during pregancy (1 = yes; 0 = no), history of premature labor PTL (1 = yes; 0 = no), history of hypertension HT (1 = yes, 0 = no), presence of intrauterine irritability UI (1 = yes, 0 = no). Use function clogit to carry out conditional regression analyses. Command: library(survival) will make it available to you in R, and then ?clogit will explain its functioning. Write a brief report on how to choose an appropriate model and express the results in terms of odds ratios and their confidence intervals.

The following three assignments deal with drug use data in Uusimaa, Finland, around 2000. The ultimate goal is to estimate the total number of drug users ($= N$), using double registration techniques. (These are also called capture-recapture techniques.) The two registers are Hospital Discharge Register (hilmo), and Criminal Report Register (riki). The data in text file drug-

users.txt have characteristics of 2,593 individuals who have been registered by at least one of the two registers as drug users.

4. Choose first those who belong to riki. You can use command subset in R. For example, if you first read in the data with : d<-read.table("clipboard", header=TRUE), you can then do: d0<-subset(d,riki==1) to get those in riki into d0. Model the probability that they also belong to hilmo using age, sex, and family status as explanatory variables. Choose the model that is, in your opinion, the best, and store the fitted values, so that they can later be used in assignment 6. Describe all this in a brief report.

5. Repeat the previous analyses to those who belong to hilmo, and model the probability that they also belong to riki.

6. Use the coefficients estimated in 4 to calculate probabilities of becoming registered in hilmo. Denote the estimates as $p_{1i}, i = 1, 2, \ldots, 2593$. Similarly, using estimates from 5, compute estimates of belonging to riki for everyone, $p_{2i}, i = 1, 2, \ldots, 2593$. The capture probabilities may vary from one drug user to another, but if the two registers can be assumed to operate independently, then the probability of becoming registered by at least one of them, equals $\phi_i = p_{1i} + p_{2i} - p_{1i}p_{2i}$ for every registered drug user in Uusimaa. Then, the so called Horvitz-Thompson estimator of the the whole population becomes $\hat{N} = \sum_{i=1}^{2593} 1/\phi_i$. How many drug users appears there to have been in Uusimaa, around 2000? Describe this analysis in a brief report, and in particular describe the estimated probabilities using histograms and statistical summaries.