Exercises for the lecture on

Statistical Methods

Prof. Dr. Vladimir Spokoiny
Vladimir.Spokoiny@wias-berlin.de
Dr. Thorsten Dickhaus
Thorsten.Dickhaus@wias-berlin.de
www.wias-berlin.de/people/dickhaus/

Humboldt-University Berlin
Department of Mathematics
Winter term 2013 / 2014

Sheet 11

Solutions are due on Monday, January 13th, 2014, 3:15pm.
Every completely and correctly solved exercise gives 4 points.

Exercises

41. **Effect modification in the context of logistic regression.**
Consider a logistic regression model with an intercept and exactly two dichotomous covariates X_1 and X_2. For every observational unit $1 \leq i \leq n$, the associated profile of covariates is therefore given by $\vec{x}_i = (x_{i1}, x_{i2})$, where $\vec{x}_i \in \{0,1\}^2$.

    (a) Illustrate the relationship between the linear predictor η_i and $\vec{x}_i$ for an arbitrary observational unit $1 \leq i \leq n$ graphically, if (i) only the main effects of the two covariates are included in the model or (ii) an additional interaction term is included into the model.

    Hint: Assume that all regression coefficients are positive.

    (b) How does the consideration / non-consideration of the interaction term influence the odds ratios (as functions of the regression coefficients)? Take the profile of covariates $(0,0)$ as reference and derive the modeled odds ratios for observing the target event depending on the covariates in both cases (i) and (ii) considered in part (a). In both cases (i) and (ii), express the univariate influence of any of the two covariates and, additionally, the joint, bivariate influence of both covariates in terms of the respective odds ratios.

42. **Conditional logistic regression.**
Stratification is a common technique in epidemiology to adjust for (dichotomous or categorical) nuisance factors, so-called "confounders". (More specifically, a confounder is a covariate which is assumed to be associated with the response, but which is itself not target of statistical inference.) Assume that this confounder can take exactly S distinct values. In a stratified analysis, all n_s observational units in stratum $1 \leq s \leq S$ (i. e., for which the confounder takes the value s) are analyzed in a *combined manner* with respect to the k covariates of interest. This requires a modified data analysis method, which is known as conditional logistic regression.

    For this, denote the i-th dichotomous response variable in the s-th stratum by Y_{si}, where $1 \leq s \leq S$ and $1 \leq i \leq n_s$ ($\sum_{s=1}^{S} n_s = n$), and the associated profile of the covariates of interest by $\vec{x}_{si} = (x_{si1}, \ldots, x_{sik})$. Making use of these quantities, we define the following linear predictors:

$$\forall 1 \leq s \leq S: \ \forall 1 \leq i \leq n_s : \eta_{si} = \beta_{0s} + \sum_{j=1}^{k} \beta_j x_{sij}$$

    This means that we have stratum-specific intercepts, while the regression coefficients for the covariates of interest are constant across the S strata. Moreover, we make the usual

(conditional) independence assumption with respect to the response variables, and we use the (canonical) logit link to map $p(\vec{x}_{si}) = \mathbb{P}_\beta(Y_{si} = 1 | \vec{X}_{si} = \vec{x}_{si})$ onto the linear predictor η_{si}.
In order to eliminate the nuisance parameters $(\beta_{0s})_{1 \leq s \leq S}$ from the likelihood function, we optimize the likelihood function conditional to the numbers $(n_{s,1})_{1 \leq s \leq S}$ of cases in stratum $1 \leq s \leq S$, which gives the method its name.

Show that, with $\beta = (\beta_1, \ldots, \beta_k)^\top$, it holds for the conditional likelihood function given $\sum_{i=1}^{n_s} Y_{si} = n_{s,1}$ in stratum $1 \leq s \leq S$:

$$L_s^{\text{cond.}}(\beta, y) = \frac{\prod_{i=1}^{n_s} \left[\exp(\sum_{j=1}^{k} \beta_j x_{sij}) \right]^{y_{si}}}{\sum_{r=1}^{\binom{n_s}{n_{s,1}}} \prod_{i=1}^{n_s} \left[\exp(\sum_{j=1}^{k} \beta_j x_{sij}) \right]^{y_{s,\pi_r(i)}}} \tag{1}$$

In equation (1), π_r with $1 \leq r \leq \binom{n_s}{n_{s,1}}$ runs over all (distinct) permutations of the n_s elements of the stratum-specific response vector $(y_{s,1}, \ldots, y_{s,n_s})^\top$.

43. **Programming exercise: ROC analysis, Exercise 9.1 in the textbook by Chap T. Le (2003)**
"Radioactive radon is an inert gas that can migrate from soil and rock and accumulate in enclosed areas such as underground mines and homes. The radioactive decay of trace amounts of uranium in Earth's crust through radium is the source of radon, or more precisely, the isotope radon-222. Radon-222 emits alpha particles; when inhaled, alpha particles rapidly diffuse across the alveolar membrane of the lung and are transported by the blood to all parts of the body. Due to the relatively high flow rate of blood in bone marrow, this may be a biologically plausible mechanism for the development of leukemia. Table E9.1 provides some data from a case-control study to investigate the association between indoor residential radon exposure and risk of childhood acute myeloid leukemia."

Make yourself familiar with the corresponding dataset (i. e., Table E9.1) which you can download freely from the URL `http://www.biostat.umn.edu/~chap/radon.html`. If you should encounter problems in downloading the file or if you do not have access to the internet, you may alternatively get the file via USB stick during the lecturer's consulting hour.

The variables in the dataset are coded as follows.

DISEASE (1: case, 2: control)
RADON (radon concentration in Bq/m^3)
SEX (1: male, 2: female)
RACE (ethnic group, 1: white, 2: black, 3: Hispanic, 4: Asian, 5: others)
DOWNS (Down's syndrome, a known risk factor for leukemia, 1: no, 2: yes)
MSMOKE (1: mother a current smoker, 2: no, 0: unknown)
MDRINK (1: mother a current alcohol drinker, 2: no, 0: unknown)
FSMOKE (as MSMOKE, but with mother replaced by father)
FDRINK (as MDRINK, but with mother replaced by father)

(a) Fit a logistic regression model to this dataset. Interpret the resulting output of your statistics software.

(b) Derive the optimum classification (i. e., diagnostic) rule with respect to leukemia for children which were not included in the sample and of which only the profile of covariates is given based on the given dataset by means of an ROC analysis.

44. **Multiple Select.** Which of the following statements are true and which are false? Please give reasons for your respective decisions (one short sentence each is sufficient).

    1. Assume that you have fitted a logistic regression model. For a new observational unit with known profile of covariates $\vec{x}_{\text{new}}$, but unknown response status, the optimum prognosis of the response is given by $\hat{y}_{\text{new}} = \mathbf{1}_{\{\hat{p}(\vec{x}_{\text{new}}) \geq 1/2\}}$.
    2. If, under a logistic regression model, all cases with identical profile of covariates are pooled and their number is taken as a new response, then a counting data structure arises and a Poisson regression model can be fitted to assess the influence of the covariates.
    3. The problem of overdispersion cannot occur in case of logistic regression, because the variance of an indicator variable only depends on the success probability.
    4. If, in a case-control study, the observational units belong to a finite population and are included into the sample by means of a pre-defined sampling strategy, then also the baseline success probability (i. e., $p(\vec{0})$) for the target event can be inferred with a logistic regression analysis.