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 10

Solutions are due on Monday, January 6th, 2014, 3:15pm.
Every completely and correctly solved exercise gives 4 points.

Exercises

37. **Multivariate central limit theorem (CLT) 3.60.**
We consider the assumptions of Theorem 3.60 and prove the result - the multivariate CLT for $\hat{\beta}(n)$. To this end, we use our general Theorem 3.7 regarding asymptotics of maximum likelihood estimators in connection with exercise 5 (exponential families fulfill all necessary regularity assumptions). It remains to derive the Fisher information of the (product) experiment (assuming fixed, given design which we suppress notationally). Show that the following assertions hold true.

    (a) The summand L_i (say) in the (joint) log-likelihood function which corresponds to the observational unit $1 \leq i \leq n$ can be written as

$$\begin{aligned} L_i := \log(l(\vartheta_i, y_i)) &= \log(a(\vartheta_i)) + \log(b(y_i)) + y_i T(\vartheta_i) \\ &= y_i\theta_i - B(\theta_i) + C(y_i), \end{aligned}$$

    where $\theta := T(\vartheta)$ and $B(\theta) := -\log(a(\vartheta))$, $\vartheta \in \Theta$, and $C(y_i) := \log(b(y_i))$.

    (b) For the first two derivatives of L_i with respect to θ_i, we get

$$\frac{\partial L_i}{\partial \theta_i} = y_i - \frac{\partial B(\theta_i)}{\partial \theta_i} \quad \text{and} \quad \frac{\partial^2 L_i}{\partial \theta_i^2} = -\frac{\partial^2 B(\theta_i)}{\partial \theta_i^2}.$$

    Exercise 5 then yields

$$\mu_i = \mathbb{E}[Y_i] = \frac{\partial B(\theta_i)}{\partial \theta_i} \quad \text{and} \quad \mathrm{Var}(Y_i) = \frac{\partial^2 B(\theta_i)}{\partial \theta_i^2}.$$

    (c) For arbitrary link function g and any $1 \leq j \leq p$, the chain rule in connection with the relation $\eta_i = g(\mu_i) = \sum_{j=1}^p \beta_j x_{ij}$ yields

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{(y_i - \mu_i)x_{ij}}{\mathrm{Var}(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}.$$

    (d) Cox and Hinkley (1974, Section 4.8) proved that, for any pair of indices $1 \leq j_i, j_2 \leq p$ and with $\beta = (\beta_1, \ldots, \beta_p)^\top$, it holds

$$\mathbb{E}_\beta\left(\frac{\partial^2 L_i}{\partial \beta_{j_1}\partial \beta_{j_2}}\right) = -\mathbb{E}_\beta\left[\left(\frac{\partial L_i}{\partial \beta_{j_1}}\right)\left(\frac{\partial L_i}{\partial \beta_{j_2}}\right)\right].$$

This result, together with exercise 5 and by making use of the stochastic independence of all response variables, yields that the (conditional) Fisher information of the product experiment is given by $I(\beta) = X^\top W X$, where W is a diagonal matrix with elements

$$w_{ii} = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \cdot [\mathrm{Var}(Y_i)]^{-1}.$$

If, moreover, g is the canonical link, we obtain $g(\mu_i) = \eta_i = T(\vartheta_i) = \theta_i$ and, consequently, the assertion of Theorem 3.60, because

$$\frac{\partial \mu_i}{\partial \eta_i} = \mathrm{Var}(Y_i).$$

38. **Saturated model.**

(a) Consider a generalized linear model in which all k covariates are dichotomous. If, per different profile of covariates, exactly one observation is made, then the saturated model is equivalent to the model which contains an intercept, all main effects and all possible interaction effects (including all possible multi-way interactions!).

(b) A slightly modified form of the assertion under (a) holds for the case that some of the covariates are categorical (with more than two possible realizations). In such cases, for deriving the saturated model one applies a dummy coding which leads to $p = n$ columns in the design matrix, where again (multi-way) interactions are taken into account. To exemplify this, consider the age- and gender-adjusted diabetes incidence data taken from Giani and Rosenbauer (1996) in Table 1. The sizes s_i, $1 \leq i \leq n$ are here given by the stratum-specific person years under risk and the responses are the stratum-specific numbers of (newly manifested) diabetes cases. The sample size equals $n = 12$.

(i) Compute the value of the maximum log-likelihood of the saturated Poisson regression model for this dataset by means of Definition 3.62.

(ii) Formulate an equivalent model by constructing a suitable design matrix with n columns, considering the three covariates "time period", "age group" and "gender", as well as an intercept and interaction terms. Show that the maximum log-likelihood value of the so-constructed model coincides with the value computed in part (i).

Hint: The model fit in part (ii) can be performed by making use of statistics software.

39. **Programming exercise: Exercise 10.1 in the textbook by Chap T. Le (2003)**
"Inflammation of the middle ear, *otitis media* (OM), is one of the most common childhood illnesses and accounts for one-third of the practice of pediatrics during the first five years of life. Understanding the natural history of otitis media is of considerable importance, due to the morbidity for children as well as concern about long-term effects on behavior, speech, and language development. In an attempt to understand that natural history, a large group of pregnant women were enrolled and their newborns were followed from birth. The response variable is the number of episodes of otitis media in the first six months (NBER), and potential factors under investigation are upper respiratory infection (URI), sibling history of otitis media (SIBHX; 1 for yes), day care, number of cigarettes consumed a day by parents (CIGS), cotinin level (CNIN) measured from the urine of the baby (a marker for exposure to cigarette smoke), and whether the baby was born in the fall season (FALL)."

Make yourself familiar with the corresponding dataset which you can download freely from the URL `http://www.biostat.umn.edu/~chap/otitis.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.

		male			female			
Time period	Age group (in years)	Diabetes cases	Person years	emp. incidence [10^{-5} py]	Diabetes cases	Person years	emp. incidence [10^{-5} py]	emp. rel. risk m/f
1973-77	0-4	100	799125	12.514	77	762120	10.103	1.24
	5-9	180	845290	21.295	150	800870	18.730	1.14
	10-14	198	799460	24.767	176	758940	23.190	1.07
1978-82	0-4	81	681035	11.894	85	650005	13.077	0.91
	5-9	204	804575	25.355	162	768265	21.087	1.20
	10-14	277	849020	32.626	224	804165	27.855	1.17

Table 1: Data for the diabetes example in exercise 38.(b)

(a) Fit a multiple Poisson regression model to these data and test the global hypothesis that none of the aforementioned covariates is associated with the response.

(b) Consider the estimated regression coefficients and their estimated standard deviations. Draw conclusions about which covariates have an important influence on the response.

(c) Are there indications for overdispersion? If yes, fit an additional overdispersion parameter and compare the results with the results obtained in part (b).

(d) Do the covariates DAYCARE and SIBHX interact in their influence on the response?

40. **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. In a generalized linear model (GLM), the choice of the link function has an influence on whether the maximum likelihood estimator for the vector of regression coefficients can be obtained in closed form or not.
2. In presence of overdispersion, the saturated model does not lead to the optimal model fit (in terms of the maximum value of the log-likelihood function).
3. The general definition of the R-square value in Definition 3.64 reduces to that given in Definition 3.21 if a multiple linear regression model with identity link as a special case of a GLM is considered.
4. If, under a Poisson regression model, no sizes s_i, $1 \leq i \leq n$ are defined and the intensity parameter ϑ_i itself is modeled as a function of the covariates, then $g = \log$ is not the canonical link anymore.