Zobecněné lineární modely

Poissonovský model

1. Pro data v souboru poisson1.txt najděte vhodný lineární regresní model popisující závislost proměnné Y na proměnné X. Danou závislost modelujte i pomocí poissonovské regrese. Oba modely porovnejte.
[Datový soubor: poisson1.txt]

2. Pro data v souboru poisson2.txt najděte vhodný lineární regresní model popisující závislost proměnné Y na proměnné X. Danou závislost modelujte i pomocí poissonovské regrese. Oba modely porovnejte.
[Datový soubor: poisson2.txt]

3. Abundace střevlíkovitých brouků v polních monokulturách závisí na mnohých faktorech prostředí. V 21 porostech pšenice proběhla observační studie, v níž byla pomocí zemních pastí sledována hojnost brouků [m^2] (abund). Do každého porostu byly umístěny snímače teploty a intenzity slunečního záření. Ze zaznamenaných dat byla spočtena průměrná denní teplota [^{0}C] (temp) a průměrná denní sluneční aktivita [W/m^2] (sun). Zajímá nás, která ze sledovaných proměnných ovlivnila hojnost střevlíkovitých brouků.
[Datový soubor: carabid.txt]

4. The following example has a count (the number of reported cancer cases per year per clinic) as the response variable, and a single continous explanatory variable (the distance from a nuclear plant to the clinic in km). The question is whether or not proximity to the reactor affects the number of cases.
[Dataset: clusters.txt]

5. Births by caesarean section are said to be more frequent in private (fee paying) hospitals as compared to non-fee paying public hospitals. Data about total annual births and the number of caesarean sections carried out were obtained from the records of 4 private hospitals and 16 public hospitals. Can we verify this statement? [Dataset: caesareans.txt]

6. A cohort of subjects, some non-smokers and others smokers, was observed for several years. The number of cases of cancer of the lung diagnosed among the different categories was recorded. Data regarding the number of years of smoking were also obtained from each individual. For each category the person-years of observation were calculated. The investigators wish to address the question of the relative risks of smoking. Person-years = The product of the number of years times the number of members of a population who have been affected by a certain condition (years of treatment with a given drug).
[Dataset: lung_cancer.txt]

Logistický a probitový model

1. Produkce kokonů (tj. útvarů obsahující vajíčka) je u pavouků ovlivněna řadou faktorů. Jedním z nich je velikost těla. V laboratorní studii byla produkce kokonů sledována u jedinců vybraného druhu pavouka s různou velikostí těla [mm] (body). Jelikož bylo možné velikost měřit pouze s přesností 0,5 mm, bylo celkem 160 jedinců (chovaných odděleně) zařazeno podle velikostní kategorie do skupin o 15 až 30 jedincích (n). Po několika dnech byl stanoven počet samic, které vyprodukovaly kokon (eggs). Nulová hypotéza, kterou budeme testovat, je následující: produkce kokonů nebyla závislá na velikosti těla samic. Pokud nulovou hypotézu zamítneme, chceme zjistit, jaký je tvar závislosti, tj. odhadnout parametry závislostí abychom mohli výsledný model použít pro predikci kladení vajíček pro velikost pavouků v intervalu 3–12 mm. [Datový soubor: spider.txt]

2. A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable. The dataset has a binary response (outcome, dependent) variable called admit. There are three predictor variables: gre, gpa and rank. We will treat the variables gre and gpa as continuous. The variable rank takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest.
   [Dataset: admission.txt]

3. In credit business, banks are interested in information whether prospective consumers will pay back their credit or not. The aim of credit-scoring is to model or predict the probability that a consumer with certain covariates is to be considered as a potential risk. The dataset consists of 1000 consumer credits from a German bank. For each consumer the binary response variable "creditability" is available. In addition, 20 covariates that are assumed to influence creditability were recorded . Find a model describing the likelihood of obtaining a loan (http://www.statistik.lmu.de/service/datenarchiv/kredit/kredit_e.html).
   [Dataset: credit_scoringn.txt]

4. In the mtcars data set, the variable vs indicates if a car has a V engine or a straight engine. We want to create a model that helps us to predict the probability of a vehicle having a V engine or a straight engine given a weight of 2100 lbs and engine displacement of 180 cubic inches. [Dataset: data(mtcars)]

5. Annual financial data are collected for bankrupt firms approximately 2 years prior to their bankruptcy and for financially sound firms at about the same time. The analysed dataset contains variables: CF/TD (cash flow/total dept), NI/TA (net income/total assets), CA/CL (current assets/current liabilities), CA/NS (current assets/net sales). Find a logit regression model and use it for classification.
   [Dataset: bankrupt_firms.txt]