Logistinen regressio
Harjoitukset 22.2.2001

3.1. Kun `penalty` aineistossa rajoitutaan tarkastelemaan vain ei-mustia syytettyjä ja ristiintaulukoidaan muuttujat `culp` ja `death`, huomataan, että arvolla `culp=1` ei ole lainkaan kuolemanrangaistuksia. Tästä aiheutuu ns. kvasiseparoituvuusongelma. Estimoidaan kuolemanrangaistuksen logitmalli, jossa selittäjinä ovat `culp` ja `serious`, missä `culp` on luokitteluasteikollinen (vrt. luennot).

```
PROC GENMOD DATA=LogReg.penalty;
  WHERE blackd=0;
  CLASS culp;
  MODEL death = culp serious / D=B;
RUN;
```

Estimoi sitten sama malli, mutta jätä aineistosta luokka `culp=1` pois. Vertaile tuloksia.

3.2. Olkoot $n_1 = 15,\ n_2 = 12$ ja $n_3 = 15$ henkilöiden lukumäärät ikäluokissa $1, 2$ ja 3 ($30-34,\ 35-39$ ja $40-44$ vuotta). Vastaavasti sairastuneiden lukumäärät eri luokissa ovat $y_1 = 2,\ y_2 = 3$ ja $y_3 = 5$. Havainnot eri luokissa ovat toisistaan riippumattomia. Oletetaan logitmalli

$$\log(\frac{\pi_i}{1-\pi_i}) = \alpha + \beta\, i,\ \ i = 1, 2, 3.$$

ja estimoidaan parametrit α ja β.

(a) Mitä on tässä tapauksessa uskottavuusfunktion maksimi?
(b) Logaritmoidun uskottavuusfunktion maksimi?

3.3. Lausu edellisen tehtävän tapauksessa

(a) täyden mallin logaritmoidun uskottavuusfunktion maksimi.
(b) devianssi tarkasteltavan malin ja täyden mallin logaritmoitujen uskottavuusfunktioiden maksimien avulla.

3.4. Selitä logistisella regressiomallilla aineistosta `chddata` mahdollisuutta sairastua sepelvaltimotautiin, kun selittäjänä on ikä. Selitä muuttujaa `chd` iällä, kun pidät ikää luokitteluasteikollisena muuttujana. Vertaile luokitteluasteikollisen selittäjän mallia malliin

$$\log(\frac{\pi_i}{1-\pi_i}) = \alpha + \beta \text{ age}.$$

3.5. Tarkastellaan ryhmitellyssä `penalty` aineistossa mallia

```
MODEL death = blackd whitvic  / D=B;
```

Mikä on tässä tapuksessa ryhmitellyn aineiston täysi malli. Laske mallin devianssi.