Logistinen regressio
Harjoitukset 8.2.2001

1.1. Oletetaan, että toisistaan riippumattomat havainnot (satunnaismuuttujat) Y_i, $i = 1, 2, \ldots, n$ noudattavat Bernoullin jakaumaa parametrein π_i (merkitään $Y_i \sim Ber(\pi_i)$) ts. satunnaismuuttujilla Y_i on tämälleen kaksi tulosvaihtoehtoa (1 ja 0) ja $P(Y_i = 1) = \pi_i$ ja $P(Y_i = 0) = 1 - \pi_i$. Oletetaan myös, että havainnot toteuttavat lineaarisen todennäköisyysmallin [Luennot malli (2.3.1)], missä $-\infty < x_i < \infty$. Laske havaintojen varianssi. Voiko se olla sama kaikilla havainnoilla?

1.2. Esitä ja perustele todennäköisyyden ja suhteellisen mahdollisuuden lausekkeet (2.5.5) ja (2.5.6) usean selittäjän mallin (2.5.7) tapauksessa.

1.3. Vertaile ohjelmien (ks. luentomonisteen s. 7)

```
PROC LOGISTIC DATA=LogReg.penalty DESCENDING;
  MODEL death=blackd whitvic serious;
RUN;
```

ja

```
PROC LOGISTIC DATA=LogReg.penalty;
  MODEL death=blackd whitvic serious;
RUN;
```

tulostuksia. LOGISTIC proseduuri mallintaa todennäköisyyden $P(Y = 0)$ ja käyttää muunnosta logit$[P(Y = 0)]$. Parametri DESCENDING muuttaa oletuksen siten, että mallinnetaan todennäköisyyttä $P(Y = 1)$, mikä on normaalikäytäntö. Selitä, mistä erot parametrien estimaateissa johtuvat.

1.4. Tulkitse edellisen tehtävän tulostuksen antamat vedonlyöntisuhteet. Miten ne lasketaan mallin regressiokertoimien estimaattien avulla?

1.5. Tarkastellaan aineistoa `chddata` (ks. `http://mtl.uta.fi/tilasto/logreg/Datat/` sekä myös tiedosto `chdinfo`). Tutki aineistoa tarkastelemalla, miten mahdollisuus sairastua sepelvaltimotautiin riippuu henkilön iästä. Jaa aineisto esimerkiksi 5 vuoden välein ikäryhmiin ja laske ikäryhmissä mahdollisuuksien suhteen (odds) estimaatti. Käytä sitten vedonlyöntisuhdetta odds:ien vertailuun.