Bweight2: Comparing log-linear models and logistic regression

```
> bweight = read.table("http://www.utstat.toronto.edu/~brunner/312f10/code_n_data/
bweight.data")
> bweight[1:5,]
   low age lwt race smoke ptl ht ui ftv  bwt
85   0  19 182    2     0   0  0  1   0 2523
86   0  33 155    3     0   0  0  0   3 2551
87   0  20 105    1     1   0  0  0   1 2557
88   0  21 108    1     1   0  0  1   2 2594
89   0  18 107    1     1   0  0  1   0 2600
>
> # Confine attention to smoking, race, low birth weight
> race <- factor(bweight$race,label=c("White","Black","Other"))
> contrasts(race)
      Black Other
White     0     0
Black     1     0
Other     0     1
> contrasts(race) <- contr.sum # Effect coding
> contrasts(race)
      [,1] [,2]
White    1    0
Black    0    1
Other   -1   -1
> smoke <- factor(bweight$smoke,label=c("No","Yes"))
> contrasts(smoke) <- contr.sum
> contrasts(smoke)
    [,1]
No     1
Yes   -1
> low <- factor(bweight$low,label=c("No","Yes"))
> contrasts(low) <- contr.sum
>
> threeD = table(smoke,race,low)
> margin.table(threeD,c(1,3,2))
, , race = White

     low
smoke No Yes
  No  40   4
  Yes 33  19

, , race = Black

     low
smoke No Yes
  No  11   5
  Yes  4   6

, , race = Other

     low
smoke No Yes
  No  35  20
  Yes  7   5
```

```
> # Conditional log-linear model with no association between
> # explanatory and response variables
> loglin1 = loglin(threeD,list(c(1,2),3))
2 iterations: deviation 2.842171e-14
> G2 = loglin1$lrt; df = loglin1$df
> G2; df; 1-pchisq(G2,df)
[1] 17.85422
[1] 5
[1] 0.003134764
> # The equivalent logistic regression model is the null model
> logregfull = glm(low ~ smoke + race + smoke:race, family=binomial)
> # low ~ smoke*race is equivalent
> summary(logregfull)

Call:
glm(formula = low ~ smoke + race + smoke:race, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3537  -0.9508  -0.4366   1.4190   2.1899

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.68896    0.20323  -3.390 0.000699 ***
smoke1       -0.52793    0.20323  -2.598 0.009384 **
race1        -0.73837    0.26668  -2.769 0.005627 **
race2         0.49746    0.31665   1.571 0.116178
smoke1:race1 -0.34733    0.26668  -1.302 0.192778
smoke1:race2 -0.06903    0.31665  -0.218 0.827425
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 216.82  on 183  degrees of freedom
AIC: 228.82

Number of Fisher Scoring iterations: 4

> anova(logregfull)
Analysis of Deviance Table

Model: binomial, link: logit

Response: low

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev
NULL                          188    234.672
smoke        1    4.867       187    229.805
race         2    9.830       185    219.975
smoke:race   2    3.157       183    216.818

> G2b = logregfull$null.deviance-logregfull$deviance
> G2b; G2
[1] 17.85422
[1] 17.85422
>
> # Connection between MLEs for the 2 kinds of model:
> # Messy for 3 and higher-D tables
>
```

```
> # Z-tests for loglinfull suggest a logistic regression model
> # without the smoke by race interaction. This is equivalent to a
> # log-linear model without the smoke by race by low interaction.
> # In general, a main effect in logistic regression corresponds to
> # an interaction between that variable and the response variable
> # in a log-linear model -- provided, of course, that the log-linear
> # model also has all interactions among explanatory variables.
> # A k-factor interaction in logistic regression corresponds to a
> # k+1-factor interaction in a log-linear model, The k+1-factor interaction
> # has all the explanatory variables in the k-factor interaction, plus
> # the response variable. Again, this is assuming  that the log-linear
> # model has all interactions among explanatory variables.
>
> # Conduct this two-df test both ways, using LR tests.
> # First with logistic regression:
>
> logregreduced = glm(low ~ smoke + race, family=binomial)
> anodev = anova(logregreduced,logregfull); anodev

Analysis of Deviance Table

Model 1: low ~ smoke + race
Model 2: low ~ smoke + race + smoke:race
  Resid. Df Resid. Dev  Df Deviance
1       185    219.975
2       183    216.818   2    3.157
> anodev[2,3]; anodev[2,4]
[1] 2
[1] 3.156937

> # Now a log-linear model. Only mu123 is missing
> loglin2 = loglin(threeD,list(c(1,2),c(1,3),c(2,3)))
5 iterations: deviation 0.08003072
> loglin2
$lrt
[1] 3.157074

$pearson
[1] 3.113864

$df
[1] 2

$margin
$margin[[1]]
[1] "smoke" "race"

$margin[[2]]
[1] "smoke" "low"

$margin[[3]]
[1] "race" "low"


> 1-pchisq(loglin2$lrt,loglin2$df)
[1] 0.2062767
> 1-pchisq(anodev[2,4],anodev[2,3])
[1] 0.2062908
> # The no-interaction logistic regression model is fine
> # [smoke race] [smoke low] [race low]
```

```
> summary(logregreduced)

Call:
glm(formula = low ~ smoke + race, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3442  -0.8862  -0.5428   1.4964   1.9939

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.5517     0.1833  -3.009  0.00262 **
smoke1       -0.5580     0.1846  -3.023  0.00251 **
race1        -0.7309     0.2490  -2.936  0.00333 **
race2         0.3532     0.2992   1.181  0.23776
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 219.97  on 185  degrees of freedom
AIC: 227.97

Number of Fisher Scoring iterations: 4

> # Why is the coefficient for smoke negative?
>
> # Test race controlling for smoke
> anodev2 = anova(logregreduced); anodev2
Analysis of Deviance Table

Model: binomial, link: logit

Response: low

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev
NULL                     188    234.672
smoke   1    4.867       187    229.805
race    2    9.830       185    219.975
> 1-pchisq(anodev2[3,2],anodev2[3,1])
[1] 0.007336125
> # Or,
> loglin3 = loglin(threeD,list(c(1,2),c(1,3)))
2 iterations: deviation 0
> G2change = loglin3$lrt-loglin2$lrt; G2change
[1] 9.829752
> dfchange = loglin3$df-loglin2$df; dfchange
[1] 2
> 1-pchisq(G2change,dfchange)
[1] 0.007336629
```

```
> # For ease of interpretation, prefer indicator dummy vars
> # when there are no interactions.
> race <- factor(bweight$race,label=c("White","Black","Other"))
> contrasts(race)
      Black Other
White     0     0
Black     1     0
Other     0     1
> smoke <- factor(bweight$smoke,label=c("No","Yes"))
> contrasts(smoke)
    Yes
No    0
Yes   1
> # Could have done: contrasts(smoke) <- contr.treatment
> # But labels were lost when we moved to effect coding
> logregreduced = glm(low ~ smoke + race, family=binomial)
> summary(logregreduced)

Call:
glm(formula = low ~ smoke + race, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3442  -0.8862  -0.5428   1.4964   1.9939

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8405     0.3529  -5.216 1.83e-07 ***
smokeYes      1.1160     0.3692   3.023  0.00251 **
raceBlack     1.0841     0.4900   2.212  0.02693 *
raceOther     1.1086     0.4003   2.769  0.00562 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 219.97  on 185  degrees of freedom
AIC: 227.97

Number of Fisher Scoring iterations: 4
```

Using the `glm` function on data that come in table format

```
> # Help says:
> # "For binomial and quasibinomial families the response can also be
> # specified as ... a two-column matrix with the columns giving the
> # numbers of successes and failures."
>
> margin.table(threeD,c(2,3,1))
, , smoke = No

       low
race    No Yes
  White 40   4
  Black 11   5
  Other 35  20

, , smoke = Yes

       low
race    No Yes
  White 33  19
  Black  4   6
  Other  7   5

> # Make a data frame from the output, and ...
> testdata <- read.table("TestFrame.txt"); testdata
  smoke  race No Yes
1    No White 40   4
2    No Black 11   5
3    No Other 35  20
4   Yes White 33  19
5   Yes Black  4   6
6   Yes Other  7   5

> Smoke <- factor(testdata$smoke); contrasts(Smoke)
    Yes
No    0
Yes   1
> Race <- factor(testdata$race,levels=c("White","Black","Other"))
> # Otherwise, alphabetical order makes Black the reference category
> contrasts(Race)
      Black Other
White     0     0
Black     1     0
Other     0     1
> # Recall we had trouble earlier controlling order of categories
> # in tables. The levels parameter will do the trick.
> LowBW <- cbind(testdata$Yes,testdata$No); LowBW
     [,1] [,2]
[1,]    4   40
[2,]    5   11
[3,]   20   35
[4,]   19   33
[5,]    6    4
[6,]    5    7
>
> # Notice order must be Yes, No!
>
> summary(glm(LowBW ~ Smoke + Race, family=binomial))
```

```
> summary(glm(LowBW ~ Smoke + Race, family=binomial))

Call:
glm(formula = LowBW ~ Smoke + Race, family = binomial)

Deviance Residuals:
       1         2         3         4         5         6
-0.93864  -0.05946   0.60978   0.59394   0.07123  -1.24205

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8405     0.3529  -5.216 1.83e-07 ***
SmokeYes      1.1160     0.3692   3.023  0.00251 **
RaceBlack     1.0841     0.4900   2.212  0.02693 *
RaceOther     1.1086     0.4003   2.769  0.00562 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 17.8542  on 5  degrees of freedom
Residual deviance:  3.1569  on 2  degrees of freedom
AIC: 31.886

Number of Fisher Scoring iterations: 4

> # Compare:
> summary(logregreduced)

Call:
glm(formula = low ~ smoke + race, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3442  -0.8862  -0.5428   1.4964   1.9939

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8405     0.3529  -5.216 1.83e-07 ***
smokeYes      1.1160     0.3692   3.023  0.00251 **
raceBlack     1.0841     0.4900   2.212  0.02693 *
raceOther     1.1086     0.4003   2.769  0.00562 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 219.97  on 185  degrees of freedom
AIC: 227.97

Number of Fisher Scoring iterations: 4

> 17.8542-3.1569
[1] 14.6973
> logregreduced$null.deviance-logregreduced$deviance
[1] 14.69729
>
```