Exposure to Dust - Logistic Regression and Search for Outliers

February 1, 2012

First of all, the dust data are loaded:

```
> library(catdata)
> data(dust)
> attach(dust)
```

First, the subsample of non-smokers is considered. A main effect logit model yields the following results:

```
> dustlogitnon1=glm(bronch ~ dust+years, family=binomial, data=dust[(dust$smoke==0),])
> summary(dustlogitnon1)

Call:
glm(formula = bronch ~ dust + years, family = binomial, data = dust[(dust$smoke ==
    0), ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0980  -0.6097  -0.4826  -0.3744   2.3608

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.157042   0.441537  -7.150 8.67e-13 ***
dust         0.005321   0.056392   0.094    0.925
years        0.053162   0.013159   4.040 5.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 282.45  on 324  degrees of freedom
Residual deviance: 264.83  on 322  degrees of freedom
AIC: 270.83

Number of Fisher Scoring iterations: 5
```

The same model as above is used without observation 1245 which can be regarded as an outlier:

```
> dustlogitnon2 <- glm(bronch ~ dust+years, family=binomial, data=dust[(dust$smoke==0)&(du
> summary(dustlogitnon2)

Call:
glm(formula = bronch ~ dust + years, family = binomial, data = dust[(dust$smoke ==
    0) & (dust$dust < 10), ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1117  -0.6149  -0.4802  -0.3730   2.3607

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.16577    0.44190  -7.164 7.84e-13 ***
dust         0.01200    0.05802   0.207    0.836
years        0.05293    0.01315   4.026 5.67e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 282.10  on 323  degrees of freedom
Residual deviance: 264.46  on 321  degrees of freedom
AIC: 270.46

Number of Fisher Scoring iterations: 5
```

The following calculations are based on the complete dataset. Therefore, main effect logit models are fitted for all observations and without observation 1246, respectively:

```
> dustlogit1 <- glm(bronch ~ dust+years+smoke, family=binomial, data=dust)
> summary(dustlogit1)

Call:
glm(formula = bronch ~ dust + years + smoke, family = binomial,
    data = dust)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3675  -0.7798  -0.5906  -0.3813   2.3022

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.047872   0.248570 -12.262  < 2e-16 ***
dust         0.091888   0.023243   3.953 7.71e-05 ***
years        0.040155   0.006206   6.470 9.78e-11 ***
smoke        0.676844   0.174380   3.881 0.000104 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1356.8  on 1245  degrees of freedom
Residual deviance: 1278.3  on 1242  degrees of freedom
AIC: 1286.3

Number of Fisher Scoring iterations: 4

> dustlogit2 <- glm(bronch ~ dust+years+smoke, family=binomial, data=dust[(dust$dust<20),]
> summary(dustlogit2)

Call:
glm(formula = bronch ~ dust + years + smoke, family = binomial,
    data = dust[(dust$dust < 20), ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2998  -0.7799  -0.5875  -0.3795   2.3043

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.061962   0.249150 -12.290  < 2e-16 ***
dust         0.099175   0.023905   4.149 3.34e-05 ***
years        0.039790   0.006213   6.404 1.51e-10 ***
smoke        0.681604   0.174525   3.905 9.40e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1356.3  on 1244  degrees of freedom
Residual deviance: 1276.3  on 1241  degrees of freedom
AIC: 1284.3

Number of Fisher Scoring iterations: 4
```