Matematiktest i børnehaveklasse

Vi skal kigge på resultaterne af en matematiktest i en børnehaveklasse. Data er et uddrag af et større datasæt, hvor vi har indskrænket til

- fire spørgsmål benævnt A,B,C,D.
- 5 drenge og 5 piger benævnt A,B,. . . ,I,J.
- 2 testtidspunkter oktober og maj.
- A,B og C består af 4 delspørgsmål, mens D har 5 delspørgsmål.

Model

Hertil bør det nævnes at vi laver en binomial fordeling for hvert spørgsmål, hver person og hver måned. eks. man har $Y_{i,A,okt}$ hvor man så køre på tværs af alle deltager, hefter har man $Y_{1,j,okt}$ med alle spørgsmål, på den måde gøres det på tværs af alt data.

Lad i angive en person, j et spørgsmål og k en måned.

- Antag at personen har sandsynlighed p_{ijk} for at svare rigtigt på hvert af delspørgsmålene og at svarene er uafhængige.
- Lad Y_{ijk} være antal rigtige svar på delspørgsmålene.

Argumenter for at Y_{ijk} er binomialfordelt med sandsynlighedsparameter p_{ijk} og antalsparameter n_j, hvor $n_A = n_B = n_C = 4$ og $n_D = 5$. Dvs $Y_{ijk} \sim bin(n_j, p_{ijk})$.

Diskuter for og imod denne model. Det følger definitionen for binomial fordeling, da der enten er et sandt eller falsk udfald, heraf binomial, og alle udfald er uafhængige. Om det så er rigtigt at antage at alle spørgsmål mm. er lige svære for alle personer er nok ikke helt fair. Det antages at alle spørgsmål i en delopgave er lige svære. ### Indlæsning af data

```
testData=read.csv("http://people.math.aau.dk/~svante/PDP3/data/pisa.csv")[,-1]
#head(testData,4)
#tail(testData,4)
testData
```

```
##    Id koen opgave resultat max maaned
## 1   A    p      A        1   4    okt
## 2   B    p      A        1   4    okt
## 3   C    p      A        1   4    okt
## 4   D    p      A        1   4    okt
## 5   E    d      A        2   4    okt
## 6   F    p      A        3   4    okt
## 7   G    d      A        1   4    okt
## 8   H    d      A        0   4    okt
## 9   I    d      A        1   4    okt
## 10  J    d      A        4   4    okt
## 11  A    p      B        2   4    okt
## 12  B    p      B        4   4    okt
## 13  C    p      B        4   4    okt
## 14  D    p      B        4   4    okt
## 15  E    d      B        4   4    okt
## 16  F    p      B        4   4    okt
## 17  G    d      B        2   4    okt
## 18  H    d      B        4   4    okt
## 19  I    d      B        2   4    okt
## 20  J    d      B        4   4    okt
## 21  A    p      C        0   4    okt
```

```
## 22  B    p    C    0  4  okt
## 23  C    p    C    0  4  okt
## 24  D    p    C    0  4  okt
## 25  E    d    C    0  4  okt
## 26  F    p    C    0  4  okt
## 27  G    d    C    0  4  okt
## 28  H    d    C    0  4  okt
## 29  I    d    C    0  4  okt
## 30  J    d    C    2  4  okt
## 31  A    p    D    2  5  okt
## 32  B    p    D    2  5  okt
## 33  C    p    D    2  5  okt
## 34  D    p    D    2  5  okt
## 35  E    d    D    0  5  okt
## 36  F    p    D    3  5  okt
## 37  G    d    D    2  5  okt
## 38  H    d    D    2  5  okt
## 39  I    d    D    2  5  okt
## 40  J    d    D    2  5  okt
## 41  A    p    A    4  4  maj
## 42  B    p    A    4  4  maj
## 43  C    p    A    4  4  maj
## 44  D    p    A    4  4  maj
## 45  E    d    A    4  4  maj
## 46  F    p    A    4  4  maj
## 47  G    d    A    4  4  maj
## 48  H    d    A    4  4  maj
## 49  I    d    A    4  4  maj
## 50  J    d    A    4  4  maj
## 51  A    p    B    2  4  maj
## 52  B    p    B    2  4  maj
## 53  C    p    B    4  4  maj
## 54  D    p    B    4  4  maj
## 55  E    d    B    4  4  maj
## 56  F    p    B    4  4  maj
## 57  G    d    B    2  4  maj
## 58  H    d    B    4  4  maj
## 59  I    d    B    3  4  maj
## 60  J    d    B    4  4  maj
## 61  A    p    C    4  4  maj
## 62  B    p    C    1  4  maj
## 63  C    p    C    4  4  maj
## 64  D    p    C    4  4  maj
## 65  E    d    C    4  4  maj
## 66  F    p    C    4  4  maj
## 67  G    d    C    2  4  maj
## 68  H    d    C    4  4  maj
## 69  I    d    C    1  4  maj
## 70  J    d    C    4  4  maj
## 71  A    p    D    4  5  maj
## 72  B    p    D    5  5  maj
## 73  C    p    D    4  5  maj
## 74  D    p    D    3  5  maj
## 75  E    d    D    4  5  maj
```

```
## 76   F    p      D         5   5    maj
## 77   G    d      D         5   5    maj
## 78   H    d      D         5   5    maj
## 79   I    d      D         1   5    maj
## 80   J    d      D         5   5    maj
```

- `resultat` er antal rigtige svar på delspørgsmålene, dvs responsen Y.
- `max` er antal delspørgsmål.
- `maaned`: testen er afholdt første gang i oktober og anden gang i maj, så man kan undersøge for progression.
- `koen`: faktoren køn har niveauerne p(ige) og d(reng).

Model med hovedvirkninger

Vi skal studerende en logisk regression med hovedvirkninger af `opgave`,`koen` og `maaned`, dvs

$$log(\frac{p_{ijk}}{1-p_{ijk}}) = koen_i + opgave_j + maaned_k$$

Denne model fittes med kommandoen

`fit=glm(cbind(resultat,max-resultat)~.,data=testData,family=binomial)`

```
fit=glm(cbind(resultat,max-resultat)~.,data=testData[,-1],family=binomial)
(summary(fit))
```

```
##
## Call:
## glm(formula = cbind(resultat, max - resultat) ~ ., family = binomial,
##     data = testData[, -1])
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.2747     0.3826   5.946 2.75e-09 ***
## koenp         0.1493     0.2734   0.546 0.585062
## opgaveB       1.0372     0.4270   2.429 0.015135 *
## opgaveC      -1.5330     0.4024  -3.810 0.000139 ***
## opgaveD      -0.5196     0.3683  -1.411 0.158275
## maanedokt    -2.5007     0.3065  -8.160 3.36e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 263.92  on 79  degrees of freedom
## Residual deviance: 144.58  on 74  degrees of freedom
## AIC: 219.89
##
## Number of Fisher Scoring iterations: 5
```

HUSK, at dette er log(odds) Det vil altså sige, at intercept op 2.3 fortæller at i opgave A, maj måned har en dreng log(odds) for at svare rigtigt på 2.3. Vi kunne for eks udregne sandsynligheden for at svare rigtigt i opgave A til at være $P(2.28) = \frac{e^2.28}{1+e^2.28} = 0.907$

- Betragt outputtet af kommandoen `summary(fit)` og konkluderer at effekten af koen er insignifikant. Koen er ikke af betydning, da vi under antagelsen H_0 siger at dregn og pige er ens, med en std. error på 0.27 vil have 59% sandsynlighed for at opservere sådan et udfald, det vil altså sige at det er højest sandsynligt at opservere dette udfald hvis der ikke er forskel på koen.

Dette begrunder at vi reducerer modellen til

```
fit0=glm(cbind(resultat,max-resultat)~ opgave+maaned,data=testData,family=binomial)
summary(fit0)
```

```
##
## Call:
## glm(formula = cbind(resultat, max - resultat) ~ opgave + maaned,
##     family = binomial, data = testData)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.3469     0.3604   6.511 7.45e-11 ***
## opgaveB       1.0359     0.4267   2.428  0.01519 *
## opgaveC      -1.5316     0.4022  -3.808  0.00014 ***
## opgaveD      -0.5190     0.3680  -1.410  0.15850
## maanedokt    -2.4980     0.3062  -8.159 3.39e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 263.92  on 79  degrees of freedom
## Residual deviance: 144.88  on 75  degrees of freedom
## AIC: 218.18
##
## Number of Fisher Scoring iterations: 5
```

Anvend outputtet af kommandoen `summary(fit0)` til at

- estimere sandsynligheden for i oktober at svare rigtigt på et delspørgsmål i opgave A. Vi bruger den linær funktion: $log(odds) = 2.3469 - 2.498 * x$ for maj: $P = \frac{e^{2.3469-2.498}}{1+e^{2.3469-2.498}} = 46.2\%$. Bemærk her at vi har en dummy variabel "x" der er 0 i maj og 1 i okt.
- estimere forbedringen i odds fra oktober til maj ift at svare rigtigt på et vilkårligt delspørgsmål Vi kan estimere odds forbedringen ud fra hældningen, her på -2.498, vi er dog interesseret i forbedringen fra okt. til maj, og -2.498 er fra maj til okt. derved skal vi blot flippe det til 2.498. dette giver altså at odds forbedres med $e^{2.498} = 12.16$. Der er altså 12.16 bedre odds i maj end okt. En væsentlig forbedring.
- angive et konfidensinterval for forbedringen

```
confint(fit0)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %     97.5 %
## (Intercept)  1.6721069  3.0879780
## opgaveB      0.2136508  1.8943783
## opgaveC     -2.3425422 -0.7607051
## opgaveD     -1.2495732  0.1975053
## maanedokt   -3.1247618 -1.9199563
```

Alternativt kan vi udregne for intercept manuelt, ved antagelse af normalfordeling: 2.3469 +/- 0.3604*1.96 = 1.641 - 3.053 Det her er i log(odds) vi vil have den i odds: $2.5\% = e^{1.641} = 5.16$ $97.5\% = e^{3.053} = 21.18$ Med værdier fra confint: $2.5\% = e^{1.672} = 5.32$ $97.5\% = e^{3.088} = 21.93$

- sammenligne odds for den sværeste opgave med odds for den letteste opgave Når vi ser på vores model, ser vi at i henhold til opgave A, der har en log(odds) på 2.35, har opgave B en på $2.35+1.04 = 3.39$, den er altså lettere end opgave, opgave C har log(odds) på $2.35-1.53 = 0.85$ og D på $2.35-0.52=1.83$. Her hvor vi har den værste log(odds) er altå for opgave C på 0.85, det må altså være den hvor en person har de dårligste odds for en success i en opgave. Derved også den sværeste opgave.