Interactions in Logistic Regression

> # UCBAAdmissions is a 3-D table: Gender by Dept by Admit
> # Same data in another format:
> # One col for Yes counts, another for No counts.
> Berkeley = read.table("http://www.utstat.toronto.edu/~brunner/312f12/code_n_data/Berkeley2.data")

> Berkeley
  Gender Dept Yes  No
1   Male    A 512 313
2  Female    A  89  19
3   Male    B 353 207
4  Female    B  17   8
5   Male    C 120 205
6  Female    C 202 391
7   Male    D 138 279
8  Female    D 131 244
9   Male    E  53 138
10 Female    E  94 299
11  Male    F  22 351
12 Female    F  24 317

> # Resp var is 2 cols. Second col is Y=1
> full = glm(cbind(No,Yes) ~ Dept*Gender,family=binomial,data=Berkeley)
> anova(full,test='Chisq')

Analysis of Deviance Table

Model: binomial, link: logit
Response: cbind(No, Yes)
Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>11</td>
<td>877.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dept</td>
<td>5</td>
<td>855.32</td>
<td>6</td>
<td>21.74 &lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>1.53</td>
<td>5</td>
<td>20.20 0.215928</td>
</tr>
<tr>
<td>Dept:Gender</td>
<td>5</td>
<td>20.20</td>
<td>0</td>
<td>0.00 0.001144 **</td>
</tr>
</tbody>
</table>
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
# Let's see what it means. Repeating some material from an earlier analysis ...

```r
noquote(gradeschool)

<table>
<thead>
<tr>
<th>Dept</th>
<th>MaleAcc</th>
<th>FemAcc</th>
<th>Chisq</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>62.1</td>
<td>82.4</td>
<td>17.25</td>
<td>3e-05</td>
</tr>
<tr>
<td>B</td>
<td>63</td>
<td>68</td>
<td>0.25</td>
<td>0.61447</td>
</tr>
<tr>
<td>C</td>
<td>36.9</td>
<td>34.1</td>
<td>0.75</td>
<td>0.38536</td>
</tr>
<tr>
<td>D</td>
<td>33.1</td>
<td>34.9</td>
<td>0.3</td>
<td>0.58515</td>
</tr>
<tr>
<td>E</td>
<td>27.7</td>
<td>23.9</td>
<td>1</td>
<td>0.31705</td>
</tr>
<tr>
<td>F</td>
<td>5.9</td>
<td>7</td>
<td>0.38</td>
<td>0.53542</td>
</tr>
</tbody>
</table>

Male = as.numeric(gradeschool[,2])
Female = as.numeric(gradeschool[,3])
cbind(Male,Female)

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62.1</td>
<td>82.4</td>
</tr>
<tr>
<td>2</td>
<td>63</td>
<td>68</td>
</tr>
<tr>
<td>3</td>
<td>36.9</td>
<td>34.1</td>
</tr>
<tr>
<td>4</td>
<td>33.1</td>
<td>34.9</td>
</tr>
<tr>
<td>5</td>
<td>27.7</td>
<td>23.9</td>
</tr>
<tr>
<td>6</td>
<td>5.9</td>
<td>7</td>
</tr>
</tbody>
</table>
```

# On the log scale, differences are logs of odds ratios.

```r
logMale = log(Male); logFemale = log(Female)
plot(rep(1:6,2),c(logMale,logFemale), pch=" ", axes=F,
     xlab="Department",ylab="Log Percent Acceptance")
axis(1,1:6,LETTERS[1:6])  # X axis
axis(2)                   # Y axis
lines(1:6,logFemale,lty=1); lines(1:6,logMale,lty=2)
points(1:6,logMale); points(1:6,logFemale,pch=19)
legend(2,2.5,legend="Female Applicants",lty=1,bty="n",pch=19)
legend(2,2.3,legend="Male Applicants",lty=2,bty="n",pch=1)
title("Berkeley Graduate Admissions by Department")
```
Berkeley Graduate Admissions by Department

Log Percent Acceptance

- Female Applicants
- Male Applicants

Department

A B C D E F
> summary(full)

Call:
glm(formula = cbind(No, Yes) ~ Dept * Gender, family = binomial,
    data = Berkeley)

Deviance Residuals:
    [1]  0  0  0  0  0  0  0  0  0  0  0  0

Coefficients:                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.5442     0.2527  -6.110 9.94e-10 ***
DeptB               0.7904     0.4977   1.588  0.11224
DeptC               2.2046     0.2672   8.252  < 2e-16 ***
DeptD               2.1662     0.2750   7.878 3.32e-15 ***
DeptE               2.7013     0.2790   9.682  < 2e-16 ***
DeptF               4.1250     0.3297  12.512  < 2e-16 ***
GenderMale         1.0521     0.2627   4.005  6.21e-05 ***
DeptB:GenderMale  -0.8321     0.5104  -1.630  0.10306
DeptC:GenderMale  -1.1770     0.2996  -3.929 8.53e-05 ***
DeptD:GenderMale  -0.9701     0.3026  -3.206  0.00135 **
DeptE:GenderMale  -1.2523     0.3303  -3.791  0.00015 ***
DeptF:GenderMale  -0.8632     0.4027  -2.144  0.03206 *

---
Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance:  8.7706e+02  on 11  degrees of freedom
Residual deviance: -8.8818e-15  on  0  degrees of freedom
AIC: 92.94

Number of Fisher Scoring iterations: 3
Categorical by Quantitative Interactions

- Parallel regression lines on the log scale mean that
- Log differences between groups are the same for each level of x.
- Odds ratios are the same for each level of x.
- Odds are in the same proportion at each level of x.
- Called a “proportional odds” model.

Log odds of passing $= \beta_0 + \beta_1 x + \beta_2 c_1 + \beta_3 c_2$

<table>
<thead>
<tr>
<th>Course</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>Odds of Passing $= e^{\beta_0} e^{\beta_1 x} e^{\beta_2 c_1} e^{\beta_3 c_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catch-up</td>
<td>1</td>
<td>0</td>
<td>$e^{\beta_0} e^{\beta_1 x} e^{\beta_2}$</td>
</tr>
<tr>
<td>Elite</td>
<td>0</td>
<td>1</td>
<td>$e^{\beta_0} e^{\beta_1 x} e^{\beta_3}$</td>
</tr>
<tr>
<td>Mainstream</td>
<td>0</td>
<td>0</td>
<td>$e^{\beta_0} e^{\beta_1 x}$</td>
</tr>
</tbody>
</table>

- Product terms represent departure from parallel lines.
- Translates to departure from proportional odds.
- To test proportional odds assumption, test regression coefficients of the product terms.

Log odds of passing $= \beta_0 + \beta_1 x + \beta_2 c_1 + \beta_3 c_2 + \beta_4 c_1 x + \beta_5 c_2 x$

<table>
<thead>
<tr>
<th>Course</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>Odds $= e^{\beta_0} e^{\beta_1 x} e^{\beta_2 c_1} e^{\beta_3 c_2} e^{\beta_4 c_1 x} e^{\beta_5 c_2 x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catch-up</td>
<td>1</td>
<td>0</td>
<td>$e^{\beta_0} e^{\beta_1 x} e^{\beta_2} e^{\beta_3} e^{\beta_4} e^{\beta_5}$</td>
</tr>
<tr>
<td>Elite</td>
<td>0</td>
<td>1</td>
<td>$e^{\beta_0} e^{\beta_1 x} e^{\beta_3} e^{\beta_5}$</td>
</tr>
<tr>
<td>Mainstream</td>
<td>0</td>
<td>0</td>
<td>$e^{\beta_0} e^{\beta_1 x}$</td>
</tr>
</tbody>
</table>

Odds ratios depend on the value of x.
> math = read.table("http://www.utstat.toronto.edu/~brunner/312f12/code_n_data/mathcat.data")
> math[1:5,]
  hsgpa hsengl hscalc course passed outcome
1  78  80 Yes Mainstrm No Failed
2  66  75 Yes Mainstrm Yes Passed
3  80  70 Yes Mainstrm Yes Passed
4  81  67 Yes Mainstrm Yes Passed
5  86  80 Yes Mainstrm Yes Passed
> attach(math) # Variable names are now available
>
> # Make dummy vars for course to be sure what's going on
> n=length(hsgpa)
> c1 = c2 = numeric(n)
> c1[course=='Catch-up'] = 1
> c2[course=='Elite'] = 1
> # table(c1,course); table(c2,course)
> c1gpa = c1*hsgpa; c2gpa = c2*hsgpa
>
> # Reduced model will have no interactions
> redmod = glm(passed ~ hsgpa+c1+c2, family=binomial)
> fullmod = glm(passed ~ hsgpa+c1+c2+c1gpa+c2gpa, family=binomial)
> anova(redmod,fullmod,test='Chisq')
Analysis of Deviance Table

Model 1: passed ~ hsgpa + c1 + c2
Model 2: passed ~ hsgpa + c1 + c2 + c1gpa + c2gpa

<table>
<thead>
<tr>
<th></th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>390</td>
<td>428.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>388</td>
<td>428.45</td>
<td>2</td>
<td>0.44679</td>
<td>0.7998</td>
</tr>
</tbody>
</table>
>
> # Can do it with factors
> contrasts(course) = contr.treatment(3,base=3)
> red = glm(passed ~ hsgpa+course, family=binomial)
> full = glm(passed ~ hsgpa+course+hsgpa:course, family=binomial)
> anova(red, full, test='Chisq')
Analysis of Deviance Table

Model 1: passed ~ hsgpa + course
Model 2: passed ~ hsgpa + course + hsgpa:course
   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     390     428.90                     
2     388     428.45  2  0.44679   0.7998

> anova(redmod, fullmod, test='Chisq') # For comparison
Analysis of Deviance Table

Model 1: passed ~ hsgpa + c1 + c2
Model 2: passed ~ hsgpa + c1 + c2 + c1gpa + c2gpa
   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     390     428.90                     
2     388     428.45  2  0.44679   0.7998

Consistent with proportional odds.
> summary(fullmod)

Call:
glm(formula = passed ~ hsgpa + c1 + c2 + c1gpa + c2gpa, family = binomial)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.4720  -0.9662   0.4454   0.8957   2.1617

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -14.2892    2.1537  -6.635 3.25e-11 ***  
hsgpa        0.1866     0.0274   6.817 9.30e-12 ***  
c1           -4.0831     9.1561  -0.446    0.656
 c2           -4.9421   10.3161  -0.479    0.632
 c1gpa        0.0360     0.1177   0.306    0.760
 c2gpa        0.0767     0.1349   0.568    0.570

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 530.66  on 393  degrees of freedom
Residual deviance: 428.45  on 388  degrees of freedom
AIC: 440.45

Number of Fisher Scoring iterations: 5

> summary(full)

Call:
glm(formula = passed ~ hsgpa + course + hsgpa:course, family = binomial)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.4720  -0.9662   0.4454   0.8957   2.1617

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -14.2892    2.1537  -6.635 3.25e-11 ***  
hsgpa        0.1866     0.0274   6.817 9.30e-12 ***  
course1      -4.0831     9.1561  -0.446    0.656
 course2      -4.9421   10.3161  -0.479    0.632
 hsgpa:course1 0.0360     0.1177   0.306    0.760
 hsgpa:course2 0.0767     0.1349   0.568    0.570

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 530.66  on 393  degrees of freedom
Residual deviance: 428.45  on 388  degrees of freedom
AIC: 440.45

Number of Fisher Scoring iterations: 5
> betahat = redmod$coefficients; betahat
> (Intercept)       hsgpa          c1          c2
>    -14.7375649   0.1922924  -1.2848883   0.9338170
>
> > gpa = 50:100
> > mainstream = betahat[1] + betahat[2]*gpa
> >
> > GPA = rep(gpa,3); Pass = c(catchup,elite,mainstream)
> > plot(GPA,Pass,pch=' ')
> > lines(gpa,catchup,lty=1)
> > lines(gpa,elite,lty=2)
> > lines(gpa,mainstream,lty=3)
> > title("Parallel Estimated Log Odds")
oddscu = exp(catchup); oddsel = exp(elite)
oddsmain = exp(mainstream)
Odds = c(oddscu,oddsel,oddsmain)
plot(GPA,Odds,pch=' ')
lines(gpa,oddscu,lty=1)
lines(gpa,oddsel,lty=2)
lines(gpa,oddsmain,lty=3)
title("Proportional Estimated Odds")