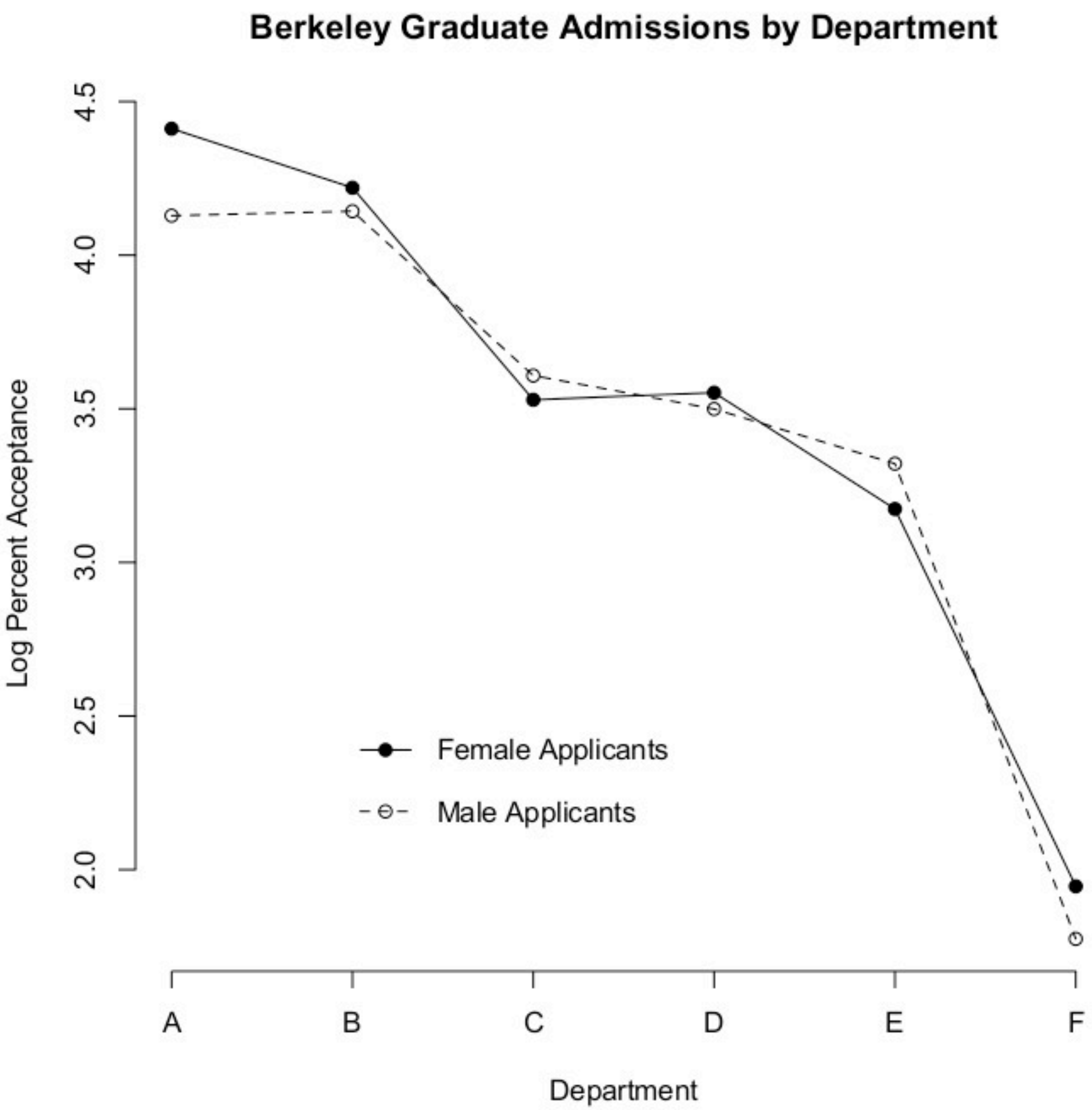
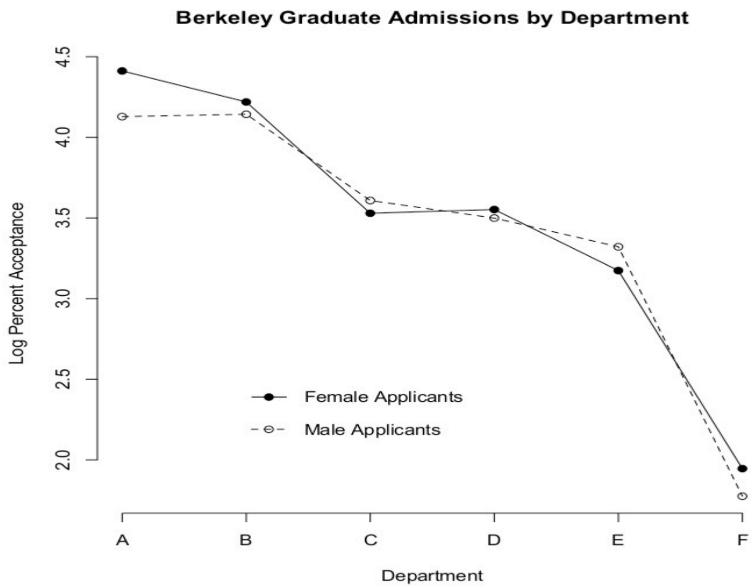
Interactions in Logistic Regression

```
> # UCBAdmissions 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)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                           11     877.06              
Dept         5   855.32         6      21.74 < 2.2e-16 ***
Gender       1     1.53         5      20.20  0.215928    
Dept:Gender  5    20.20         0       0.00  0.001144 ** 
---
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 ...
> noquote(gradschool)
     Dept %MaleAcc %FemAcc Chisq p-value
[1,] A    62.1     82.4    17.25 3e-05
[2,] B    63       68      0.25  0.61447
[3,] C    36.9     34.1    0.75  0.38536
[4,] D    33.1     34.9    0.3   0.58515
[5,] E    27.7     23.9    1     0.31705
[6,] F    5.9      7       0.38  0.53542
> Male   = as.numeric(gradschool[,2])
> Female = as.numeric(gradschool[,3])
> cbind(Male,Female)
     Male Female
[1,] 62.1   82.4
[2,] 63.0   68.0
[3,] 36.9   34.1
[4,] 33.1   34.9
[5,] 27.7   23.9
[6,]  5.9    7.0

> # On the log scale, differences are logs of odds ratios.
> # Non-parallel means the odds ratio DEPENDS
> 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
4.5
4.0
3.5
3.0
2.5
2.0
Female Applicants
Male Applicants
A
B
C
D
E
F
Department




```
> 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 ‘***’ 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.

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

Course	c_1	c_2	**Odds of Passing** $= e^{\beta_0} e^{\beta_1 x} e^{\beta_2 c_1} e^{\beta_3 c_2}$
Catch-up	1	0	$e^{\beta_0} e^{\beta_1 x}\, e^{\beta_2}$
Elite	0	1	$e^{\beta_0} e^{\beta_1 x}\, e^{\beta_3}$
Mainstream	0	0	$e^{\beta_0} e^{\beta_1 x}$

- 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.

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

Course	c_1	c_2	**Odds** $= e^{\beta_0} e^{\beta_1 x} e^{\beta_2 c_1} e^{\beta_3 c_2} e^{\beta_4 c_1 x} e^{\beta_5 c_2 x}$
Catch-up	1	0	$e^{\beta_0} e^{\beta_1 x}\, e^{\beta_2} e^{\beta_4 x}$
Elite	0	1	$e^{\beta_0} e^{\beta_1 x}\, e^{\beta_3} e^{\beta_5 x}$
Mainstream	0	0	$e^{\beta_0} e^{\beta_1 x}$

Odds ratios depend on the value of x.

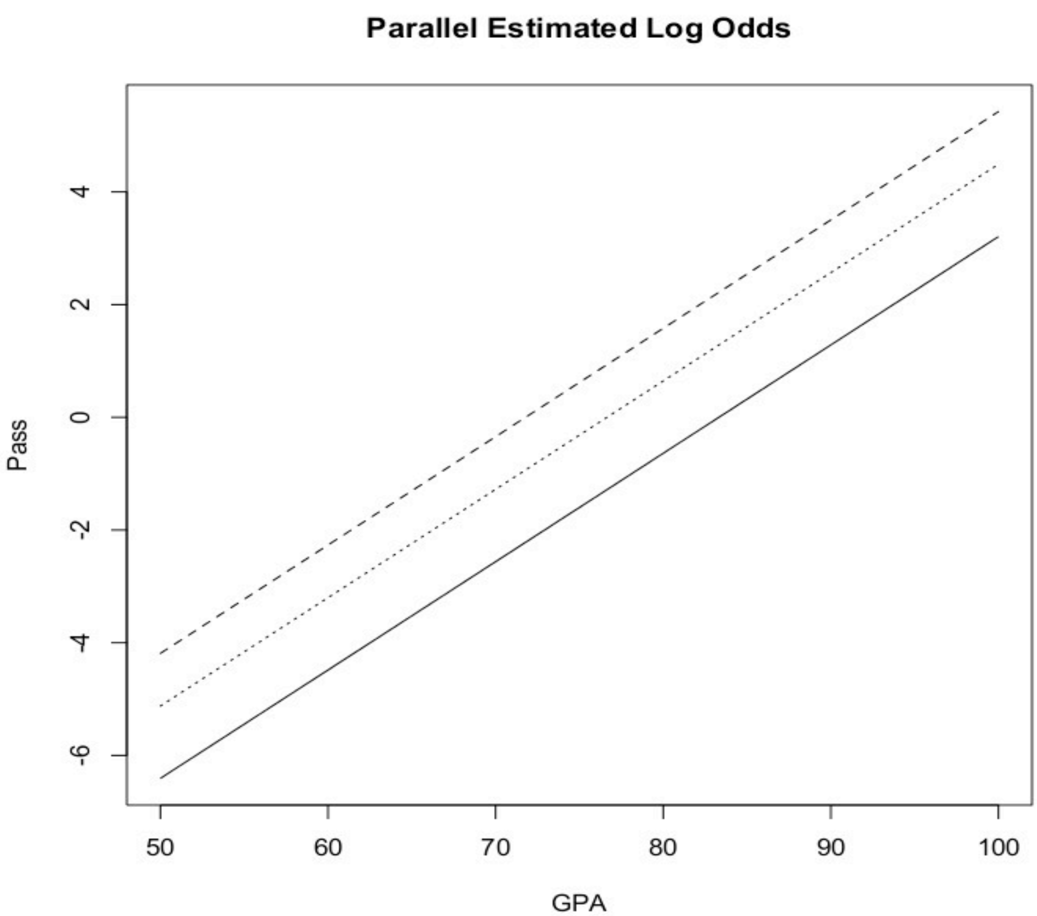
```
> math = read.table("http://www.utstat.toronto.edu/~brunner/312f12
                              /code_n_data/mathcat.data")
> math[1:5,]
  hsgpa hsengl hscalc   course passed outcome
1  78.0     80    Yes Mainstrm     No  Failed
2  66.0     75    Yes Mainstrm    Yes  Passed
3  80.2     70    Yes Mainstrm    Yes  Passed
4  81.7     67    Yes Mainstrm    Yes  Passed
5  86.8     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
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       390     428.90
2       388     428.45  2  0.44679   0.7998
>

> # 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) -14.28923    2.15367  -6.635 3.25e-11 ***
hsgpa         0.18658    0.02737   6.817 9.30e-12 ***
c1           -4.08308    9.15612  -0.446    0.656
c2           -4.94207   10.31611  -0.479    0.632
c1gpa         0.03600    0.11773   0.306    0.760
c2gpa         0.07668    0.13492   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)   -14.28923    2.15367  -6.635 3.25e-11 ***
hsgpa           0.18658    0.02737   6.817 9.30e-12 ***
course1        -4.08308    9.15612  -0.446    0.656
course2        -4.94207   10.31611  -0.479    0.632
hsgpa:course1   0.03600    0.11773   0.306    0.760
hsgpa:course2   0.07668    0.13492   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
> catchup    = betahat[1]+betahat[3] + betahat[2]*gpa
> elite      = betahat[1]+betahat[4] + betahat[2]*gpa
> 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")
```