Logistic Regression with R: Example One

```
> 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
> length(hsgpa)
[1] 394
>
> # First, some simple examples to illustrate the methods
> # Two continuous explanatory variables
> model1 = glm(passed ~ hsgpa + hsengl, family=binomial)
> summary(model1)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5577  -0.9833   0.4340   0.9126   2.2883

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.69568    2.00683  -7.323 2.43e-13 ***
hsgpa         0.22982    0.02955   7.776 7.47e-15 ***
hsengl       -0.04020    0.01709  -2.352   0.0187 *
---
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: 437.69  on 391  degrees of freedom
AIC: 443.69

Number of Fisher Scoring iterations: 4

> betahat1 =  model1$coefficients; betahat1
 (Intercept)        hsgpa       hsengl
-14.69567812   0.22982332  -0.04020062
>
> # For a constant value of mark in HS English, for every one-point increase
> # in HS GPA, estimated odds of passing are multiplied by ...
> exp(betahat1[2])
   hsgpa
1.258378
```

Deviance = -2[L_M - L_S] (p. 85)

Where L_M is the maximum log likelihood of the model, and L_S is the maximum log likelihood of an "ideal" model that fits as well as possible. The greater the deviance, the worse the model fits compared to the "best case."

Akaike information criterion: AIC = 2p + Deviance,

where p = number of model parameters

```
>
> # Deviance = -2LL + c
> # Constant will be discussed later.
> # But recall that the likelihood ratio test statistic is the
> # DIFFERENCE between two -2LL values, so
> # G-squared = Deviance(Reduced)-Deviance(Full)
>
> # Test both explanatory variables at once
> # Null deviance is deviance of a model with just the intercept.
> model1$deviance
[1] 437.6855
> model1$null.deviance
[1] 530.6559
> # G-squared = Deviance(Reduced)-Deviance(Full)
> # df = difference in number of betas
> G2 = model1$null.deviance-model1$deviance; G2
[1] 92.97039
> 1-pchisq(G2,df=1)
[1] 0
>
> a1 = anova(model1); a1
Analysis of Deviance Table

Model: binomial, link: logit

Response: passed

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev
NULL                     393     530.66
hsgpa   1   87.221       392     443.43
hsengl  1    5.749       391     437.69
> # a1 is a matrix
> a1[1,4] - a1[2,4]
[1] 87.22114
> anova(model1,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: passed

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     393     530.66
hsgpa   1   87.221       392     443.43   <2e-16 ***
hsengl  1    5.749       391     437.69   0.0165 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # For LR test of hsengl controlling for hagpa
> # Compare Z = -2.352,  p =   0.0187
```

```
>
> # Estimate the probability of passing for a student with
> # HSGPA = 80 and HS English = 75
```

$$\pi = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}$$

```
>
> x = c(1,80,75); xb = sum(x*model1$coefficients)
> phat = exp(xb)/(1+exp(xb)); phat
[1] 0.8042151


>
> ############ Categorical explanatory variables  ############
> # Are represented by dummy variables.
> # First an example from earlier.
>
> coursepassed = table(course,passed); coursepassed
          passed
course      No Yes
  Catch-up  27   8
  Elite      7  24
  Mainstrm 124 204
> addmargins(coursepassed,c(1,2)) # See marginal totals
          passed
course      No Yes Sum
  Catch-up  27   8  35
  Elite      7  24  31
  Mainstrm 124 204 328
  Sum      158 236 394
> prop.table(coursepassed,1) # See proportions of row totals
          passed
course            No       Yes
  Catch-up 0.7714286 0.2285714
  Elite    0.2258065 0.7741935
  Mainstrm 0.3780488 0.6219512
>
> # Test independence, first with a Pearson X^2
> cp = chisq.test(coursepassed); cp

      Pearson's Chi-squared test

data:  coursepassed
X-squared = 24.6745, df = 2, p-value = 4.385e-06

>
>
> # Now LR test
```

$$G^2 = 2\sum_{j=1}^{c} n_j \log\left(\frac{n_j}{\widehat{\mu}_j}\right)$$

```
> muhat = cp$expected; nij = coursepassed
> G2 = 2 * sum( nij * log(nij/muhat) ); G2
[1] 24.91574
```

```
> muhat = cp$expected; nij = coursepassed
> G2 = 2 * sum( nij * log(nij/muhat) ); G2
[1] 24.91574

> # Now with logistic regression and dummy variables
> is.factor(course) # Is course already a factor?
[1] TRUE
> contrasts(course) # Reference cat should be alphabetically first
         Elite Mainstrm
Catch-up     0        0
Elite        1        0
Mainstrm     0        1
> # Want Mainstream to be the reference category
> contrasts(course) = contr.treatment(3,base=3)
> contrasts(course)
         1 2
Catch-up 1 0
Elite    0 1
Mainstrm 0 0
>
> model2 = glm(passed ~ course, family=binomial); summary(model2)

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

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7251  -1.3948   0.9746   0.9746   1.7181  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   0.4978     0.1139   4.372 1.23e-05 ***
course1      -1.7142     0.4183  -4.098 4.17e-05 ***
course2       0.7343     0.4444   1.652   0.0985 .  
---
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: 505.74  on 391  degrees of freedom
AIC: 511.74

Number of Fisher Scoring iterations: 4

> anova(model2) # Both dummy variables are entered at once bec. course is a factor.
Analysis of Deviance Table

Model: binomial, link: logit

Response: passed

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev
NULL                     393     530.66
course  2   24.916       391     505.74
> # Compare G^2 = 24.91574 from the LR test of independence.
```

```
>
> # The estimated odds of passing are __ times as great for students in
> # the catch-up course, compared to students in the mainstream course.
> model2$coefficients
(Intercept)     course1     course2
  0.4978384  -1.7142338   0.7343053
> exp(model2$coefficients[2])
  course1
0.1801017
>
> # Get that number from the contingency table
> addmargins(coursepassed,c(1,2))
          passed
course      No Yes Sum
  Catch-up  27   8  35
  Elite      7  24  31
  Mainstrm 124 204 328
  Sum      158 236 394
> pr = prop.table(coursepassed,1); pr # Estimated conditional probabilities
          passed
course            No       Yes
  Catch-up 0.7714286 0.2285714
  Elite    0.2258065 0.7741935
  Mainstrm 0.3780488 0.6219512

> odds1 = pr[1,2]/(1-pr[1,2]); odds1
[1] 0.2962963
> odds3 = pr[3,2]/(1-pr[3,2]); odds3
[1] 1.645161
> odds1/odds3
[1] 0.1801017
> exp(model2$coefficients[2])
  course1
0.1801017
```

```
>
>
> ############### Now a more realistic analysis ####################
>
> model3 = glm(passed ~ course + hsgpa + hsengl, family=binomial)
> summary(model3)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5404  -0.9852   0.4110   0.8820   2.2109

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.18265    2.06382  -6.872 6.33e-12 ***
course1      -1.29137    0.45190  -2.858  0.00427 **
course2       0.75847    0.49308   1.538  0.12399
hsgpa         0.21939    0.02988   7.342 2.10e-13 ***
hsengl       -0.03534    0.01766  -2.001  0.04539 *
---
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: 424.76  on 389  degrees of freedom
AIC: 434.76

Number of Fisher Scoring iterations: 4

> anova(model3,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: passed

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     393     530.66
course  2   24.916       391     505.74 3.887e-06 ***
hsgpa   1   76.844       390     428.90 < 2.2e-16 ***
hsengl  1    4.132       389     424.76   0.04209 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Interpret all the tests
```

```
>
> # How about whether they took HS Calculus?
> model4 = update(model3,~ . + hscalc); summary(model4)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5517  -0.9811   0.4059   0.8716   2.2061

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.42813    2.20154  -7.008 2.42e-12 ***
course1      -0.88042    0.48834  -1.803   0.0714 .
course2       0.79966    0.50023   1.599   0.1099
hsgpa         0.22036    0.03003   7.337 2.19e-13 ***
hsengl       -0.03619    0.01776  -2.038   0.0416 *
hscalcYes     1.25718    0.67282   1.869   0.0617 .
---
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: 420.90  on 388  degrees of freedom
AIC: 432.9

Number of Fisher Scoring iterations: 4

>
> # Test course controlling for others
> notcourse = glm(passed ~ hsgpa + hsengl + hscalc , family = binomial)
> anova(notcourse,model4,test="Chisq")
Analysis of Deviance Table

Model 1: passed ~ hsgpa + hsengl + hscalc
Model 2: passed ~ course + hsgpa + hsengl + hscalc
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       390     427.75
2       388     420.90  2   6.8575   0.03243 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>
> # I like Model 3.
```

```
>
> # I like Model 3. Answer the following questions based on Model 3.
>
> # Controlling for High School english mark and High School GPA,
> # the estimated odds of passing are ___ times as great for students in the
> # Elite course, compared to students in the Catch-up course.
>
> betahat3 = model3$coefficients; betahat3
 (Intercept)      course1      course2        hsgpa       hsengl
-14.18264539  -1.29136575   0.75846785   0.21939002  -0.03533871
> exp(betahat3[3])/exp(betahat3[2])
 course2
7.766609
>
> # What is the estimated probability of passing for a student
> # in the mainstream course with 90% in HS English and a HS GPA of 80%?
>
> x = c(1,0,0,80,90); xb = sum(x*model3$coefficients)
> phat = exp(xb)/(1+exp(xb)); phat
[1] 0.54688
>
> # What if the student had 50% in HS English?
> x = c(1,0,0,80,50); xb = sum(x*model3$coefficients)
> phat = exp(xb)/(1+exp(xb)); phat
[1] 0.8322448
>
> # What if the student had -40 in HS English?
> x = c(1,0,0,80,-40); xb = sum(x*model3$coefficients)
> phat = exp(xb)/(1+exp(xb)); phat
[1] 0.9916913
>
>
```

A confidence interval would be nice.