Stats for Clinical Trials, Math 150
Jo Hardin
Logistic Regression example:
interaction & stepwise regression

Interaction

Consider data is from the Heart and Estrogen/Progestin Study (HERS), a clinical trial of hormone therapy for prevention of recurrent heart attacks and deaths among 2,763 post-menopausal women with existing coronary heart disease (Hulley et al., 1998). The HERS data is available at: `http://www.epibiostat.ucsf.edu/biostat/vgsm/data/hersdata.txt`, and it is described in **Regression Methods in Biostatistics**, page 30; variable descriptions are also given on the book website `http://www.epibiostat.ucsf.edu/biostat/vgsm/data/hersdata.codebook.txt`.

For now, we will try to predict whether the individuals had a medical condition, `medcond`. We will use the variables `age`, `weight`, `diabetes` and `drinkany`

```
HERS <- read.table("~/Dropbox/teaching/math150/HERS.csv",
                   sep=",", header=T, na.strings=".")
attach(HERS)
summary(glm(medcond ~ age, family="binomial"))$coef

##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -1.60404454 0.400644718 -4.003658 6.237044e-05
## age          0.01619155 0.005965348  2.714267 6.642259e-03

summary(glm(medcond ~ age + weight, family="binomial"))$coef

##                 Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -2.169846602 0.496466231 -4.370582 1.239155e-05
## age          0.018926204 0.006132171  3.086379 2.026105e-03
## weight       0.005279148 0.002742218  1.925138 5.421212e-02

summary(glm(medcond ~ age+diabetes, family="binomial"))$coef

##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -1.89085342 0.407555871 -4.639495 3.492616e-06
## age          0.01848156 0.006027953  3.065977 2.169602e-03
## diabetes     0.48714064 0.088177630  5.524538 3.303543e-08

summary(glm(medcond ~ age*diabetes, family="binomial"))$coef

##                 Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept)  -2.51762428 0.478275141 -5.263966 1.409802e-07
## age           0.02780399 0.007072117  3.931495 8.441917e-05
## diabetes      2.83494074 0.913870625  3.102125 1.921369e-03
## age:diabetes -0.03540210 0.013718549 -2.580601 9.862861e-03

summary(glm(medcond ~ age*drinkany, family="binomial"))$coef

##                  Estimate  Std. Error   z value   Pr(>|z|)
## (Intercept)  -0.990678407 0.511047367 -1.938526 0.05255913
## age           0.008847923 0.007588988  1.165890 0.24365904
## drinkany     -1.439698282 0.831289196 -1.731886 0.08329383
## age:drinkany  0.016793128 0.012391965  1.355163 0.17536577
```

Forward Model Building

Manual

Starting with age in the model (possibly age is an important covariate that needs to be in the model), I'll add different parameters (and interaction terms) to see if the model can be improved with more variables.

```
summary(glm(medcond ~ age, family="binomial"))$coef

##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -1.60404454 0.400644718 -4.003658 6.237044e-05
## age          0.01619155 0.005965348  2.714267 6.642259e-03

summary(glm(medcond ~ age + weight, family="binomial"))$coef

##                 Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -2.169846602 0.496466231 -4.370582 1.239155e-05
## age          0.018926204 0.006132171  3.086379 2.026105e-03
## weight       0.005279148 0.002742218  1.925138 5.421212e-02

summary(glm(medcond ~ age+ diabetes, family="binomial"))$coef

##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -1.89085342 0.407555871 -4.639495 3.492616e-06
## age          0.01848156 0.006027953  3.065977 2.169602e-03
## diabetes     0.48714064 0.088177630  5.524538 3.303543e-08

summary(glm(medcond ~ age*diabetes, family="binomial"))$coef

##                 Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept)  -2.51762428 0.478275141 -5.263966 1.409802e-07
## age           0.02780399 0.007072117  3.931495 8.441917e-05
## diabetes      2.83494074 0.913870625  3.102125 1.921369e-03
## age:diabetes -0.03540210 0.013718549 -2.580601 9.862861e-03
```

"Automatic" Drop-in-deviance

```
HERS.all = HERS[complete.cases(HERS),]   # remove all rows with any NA values
fmod1 = glm(medcond ~ 1, data=HERS.all, family="binomial")  # only the intercept
add1(fmod1, ~ age + nonwhite + smoking + drinkany + exercise + weight +
       BMI + HDL + LDL + diabetes, data=HERS.all, test="Chisq")   # check the variables listed

## Single term additions
##
## Model:
## medcond ~ 1
##          Df Deviance    AIC     LRT  Pr(>Chi)
## <none>        3383.8 3385.8
## age       1   3376.5 3380.5  7.2364  0.007144 **
## nonwhite  1   3383.7 3387.7  0.0579  0.809883
## smoking   1   3383.7 3387.7  0.0496  0.823792
## drinkany  1   3372.1 3376.1 11.6471  0.000643 ***
## exercise  1   3373.7 3377.7 10.0456  0.001527 **
## weight    1   3381.4 3385.4  2.4108  0.120500
## BMI       1   3377.5 3381.5  6.2533  0.012396 *
## HDL       1   3383.6 3387.6  0.1755  0.675232
## LDL       1   3377.1 3381.1  6.7225  0.009520 **
## diabetes  1   3354.7 3358.7 29.0868 6.921e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

fmod2 = glm(medcond ~ diabetes, data=HERS.all, family="binomial")
add1(fmod2, ~ age + nonwhite + smoking + drinkany + exercise + weight +
       BMI + HDL + LDL + diabetes, data=HERS.all, test="Chisq")   # check the variables listed

## Single term additions
##
## Model:
## medcond ~ diabetes
##          Df Deviance    AIC    LRT Pr(>Chi)
## <none>        3354.7 3358.7
## age       1   3345.4 3351.4 9.3350 0.002248 **
## nonwhite  1   3354.5 3360.5 0.1786 0.672595
## smoking   1   3354.6 3360.6 0.0632 0.801449
## drinkany  1   3348.2 3354.2 6.4411 0.011151 *
## exercise  1   3347.3 3353.3 7.3664 0.006645 **
## weight    1   3354.7 3360.7 0.0247 0.875050
## BMI       1   3353.6 3359.6 1.1182 0.290302
## HDL       1   3353.0 3359.0 1.6596 0.197656
## LDL       1   3348.6 3354.6 6.0782 0.013686 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

fmod3 = glm(medcond ~ diabetes + age, data=HERS.all, family="binomial")
add1(fmod3, ~ (age + diabetes)^2 + nonwhite + smoking + drinkany + age + weight +
       BMI + HDL + LDL, data=HERS.all, test="Chisq")   # check the variables listed
```

```
## Single term additions
##
## Model:
## medcond ~ diabetes + age
##              Df Deviance    AIC    LRT Pr(>Chi)
## <none>            3345.4 3351.4
## nonwhite      1   3345.3 3353.3 0.0073  0.93181
## smoking       1   3344.6 3352.6 0.7991  0.37137
## drinkany      1   3339.8 3347.8 5.5599  0.01838 *
## weight        1   3344.7 3352.7 0.6517  0.41951
## BMI           1   3343.1 3351.1 2.2939  0.12989
## HDL           1   3344.4 3352.4 0.9177  0.33808
## LDL           1   3339.9 3347.9 5.4086  0.02004 *
## age:diabetes  1   3339.8 3347.8 5.5430  0.01855 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fmod4 = glm(medcond ~ diabetes + age + drinkany, data=HERS.all, family="binomial")
add1(fmod4, ~ (age + diabetes + drinkany)^2 + nonwhite + smoking + +exercise+LDL +
       weight + BMI + HDL, data=HERS.all, test="Chisq")   # check the variables listed
```

```
## Single term additions
##
## Model:
## medcond ~ diabetes + age + drinkany
##                   Df Deviance    AIC    LRT Pr(>Chi)
## <none>                 3339.8 3347.8
## nonwhite           1   3339.6 3349.6 0.1666 0.683123
## smoking            1   3339.0 3349.0 0.7450 0.388052
## exercise           1   3332.4 3342.4 7.3752 0.006613 **
## LDL                1   3333.8 3343.8 5.9567 0.014661 *
## weight             1   3339.2 3349.2 0.5967 0.439825
## BMI                1   3337.8 3347.8 2.0085 0.156419
## HDL                1   3338.1 3348.1 1.7178 0.189981
## age:diabetes       1   3334.7 3344.7 5.0826 0.024167 *
## age:drinkany       1   3339.3 3349.3 0.4835 0.486862
## diabetes:drinkany  1   3339.1 3349.1 0.6576 0.417418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fmod5 = glm(medcond ~ diabetes + age + drinkany + exercise, data=HERS.all, family="binomial")
add1(fmod5, ~ nonwhite + smoking + (age + drinkany+exercise+diabetes)^2 + weight +
       BMI + HDL + LDL, data=HERS.all, test="Chisq")   # check the variables listed
```

```
## Single term additions
##
## Model:
## medcond ~ diabetes + age + drinkany + exercise
##                   Df Deviance    AIC    LRT Pr(>Chi)
## <none>                 3332.4 3342.4
## nonwhite           1   3332.2 3344.2 0.2458 0.620028
## smoking            1   3332.2 3344.2 0.2269 0.633823
```

```
## weight              1   3332.2 3344.2 0.1717 0.678625
## BMI                 1   3331.4 3343.4 1.0357 0.308826
## HDL                 1   3330.5 3342.5 1.8781 0.170550
## LDL                 1   3325.8 3337.8 6.6595 0.009863 **
## age:drinkany        1   3331.9 3343.9 0.4949 0.481770
## age:exercise        1   3332.4 3344.4 0.0067 0.934961
## age:diabetes        1   3327.4 3339.4 5.0298 0.024915 *
## drinkany:exercise   1   3332.0 3344.0 0.4104 0.521756
## drinkany:diabetes   1   3331.8 3343.8 0.5690 0.450663
## exercise:diabetes   1   3330.3 3342.3 2.0682 0.150397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

fmod6 = glm(medcond ~ diabetes + age + drinkany + exercise+LDL, data=HERS.all, family="binomial")
add1(fmod6, ~ nonwhite + smoking + (age + drinkany+exercise+LDL+diabetes)^2 + weight + BMI+
       HDL, data=HERS.all, test="Chisq")    # check the variables listed

## Single term additions
##
## Model:
## medcond ~ diabetes + age + drinkany + exercise + LDL
##                    Df Deviance    AIC    LRT Pr(>Chi)
## <none>                  3325.8 3337.8
## nonwhite            1   3325.6 3339.6 0.1384  0.70984
## smoking             1   3325.5 3339.5 0.2714  0.60240
## weight              1   3325.5 3339.5 0.2841  0.59403
## BMI                 1   3324.4 3338.4 1.3548  0.24443
## HDL                 1   3323.9 3337.9 1.8567  0.17301
## age:drinkany        1   3325.3 3339.3 0.4241  0.51490
## age:exercise        1   3325.7 3339.7 0.0138  0.90642
## age:LDL             1   3325.3 3339.3 0.4373  0.50842
## age:diabetes        1   3321.0 3335.0 4.7846  0.02871 *
## drinkany:exercise   1   3325.4 3339.4 0.3609  0.54803
## drinkany:LDL        1   3325.4 3339.4 0.3582  0.54950
## drinkany:diabetes   1   3325.2 3339.2 0.5888  0.44287
## exercise:LDL        1   3325.8 3339.8 0.0041  0.94914
## exercise:diabetes   1   3323.8 3337.8 1.9377  0.16392
## LDL:diabetes        1   3324.4 3338.4 1.3420  0.24667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

fmod7 = glm(medcond ~ diabetes*age + drinkany + exercise+LDL, data=HERS.all, family="binomial")
add1(fmod7, ~ nonwhite + smoking + (age + drinkany+exercise+LDL+diabetes)^2 + weight + BMI+
       HDL, data=HERS.all, test="Chisq")    # check the variables listed

## Single term additions
##
## Model:
## medcond ~ diabetes * age + drinkany + exercise + LDL
##                    Df Deviance    AIC     LRT Pr(>Chi)
## <none>                  3321.0 3335.0
## nonwhite            1   3320.8 3336.8 0.21626   0.6419
```

```
## smoking            1   3320.5 3336.5 0.46595   0.4949
## weight             1   3320.7 3336.7 0.27759   0.5983
## BMI                1   3319.7 3335.7 1.28869   0.2563
## HDL                1   3319.2 3335.2 1.82308   0.1769
## age:drinkany       1   3320.9 3336.9 0.05634   0.8124
## age:exercise       1   3321.0 3337.0 0.00102   0.9745
## age:LDL            1   3320.6 3336.6 0.34719   0.5557
## drinkany:exercise  1   3320.5 3336.5 0.46352   0.4960
## drinkany:LDL       1   3320.7 3336.7 0.27462   0.6003
## drinkany:diabetes  1   3320.4 3336.4 0.60596   0.4363
## exercise:LDL       1   3321.0 3337.0 0.01660   0.8975
## exercise:diabetes  1   3319.2 3335.2 1.80984   0.1785
## LDL:diabetes       1   3319.7 3335.7 1.23156   0.2671
```

No more variables are significant when added to the model. So the final model is given as:

```
summary(fmod7)$coef
```

```
##                   Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept)  -1.812683836 0.535281530 -3.386412 0.0007081296
## diabetes      2.549948179 0.964844077  2.642860 0.0082208928
## age           0.025851499 0.007357548  3.513603 0.0004420734
## drinkany     -0.201325854 0.086830107 -2.318618 0.0204157554
## exercise     -0.241959997 0.085809548 -2.819733 0.0048063651
## LDL          -0.002811455 0.001117971 -2.514783 0.0119105785
## diabetes:age -0.031664428 0.014480615 -2.186677 0.0287661096
```

Automatic **AIC**

AIC $= -2 \ln$ likelihood $+ 2p$. A small AIC is good because a big likelihood is good. However, the likelihood will continue to grow with more parameters, so the AIC is penalized by adding $2p$ (where p is the number of parameters we estimate in the model.)

```
fmod1.aic = glm(medcond ~ 1, data=HERS.all, family="binomial")  # only the intercept
# check the variables listed
step(fmod1.aic, ~ age + nonwhite + smoking + drinkany + exercise + weight +
       BMI + HDL + LDL + diabetes, data=HERS.all, direction="forward", k=2)
```

```
## Start:  AIC=3385.77
## medcond ~ 1
##
##             Df Deviance    AIC
## + diabetes   1   3354.7 3358.7
## + drinkany   1   3372.1 3376.1
## + exercise   1   3373.7 3377.7
## + age        1   3376.5 3380.5
## + LDL        1   3377.1 3381.1
## + BMI        1   3377.5 3381.5
## + weight     1   3381.4 3385.4
## <none>           3383.8 3385.8
## + HDL        1   3383.6 3387.6
```

```
## + nonwhite  1   3383.7 3387.7
## + smoking   1   3383.7 3387.7
##
## Step:  AIC=3358.69
## medcond ~ diabetes
##
##            Df Deviance    AIC
## + age       1   3345.4 3351.4
## + exercise  1   3347.3 3353.3
## + drinkany  1   3348.2 3354.2
## + LDL       1   3348.6 3354.6
## <none>          3354.7 3358.7
## + HDL       1   3353.0 3359.0
## + BMI       1   3353.6 3359.6
## + nonwhite  1   3354.5 3360.5
## + smoking   1   3354.6 3360.6
## + weight    1   3354.7 3360.7
##
## Step:  AIC=3351.35
## medcond ~ diabetes + age
##
##            Df Deviance    AIC
## + exercise  1   3337.7 3345.7
## + drinkany  1   3339.8 3347.8
## + LDL       1   3339.9 3347.9
## + BMI       1   3343.1 3351.1
## <none>          3345.4 3351.4
## + HDL       1   3344.4 3352.4
## + smoking   1   3344.6 3352.6
## + weight    1   3344.7 3352.7
## + nonwhite  1   3345.3 3353.3
##
## Step:  AIC=3345.72
## medcond ~ diabetes + age + exercise
##
##            Df Deviance    AIC
## + LDL       1   3331.6 3341.6
## + drinkany  1   3332.4 3342.4
## <none>          3337.7 3345.7
## + BMI       1   3336.5 3346.5
## + HDL       1   3336.7 3346.7
## + smoking   1   3337.5 3347.5
## + weight    1   3337.5 3347.5
## + nonwhite  1   3337.7 3347.7
##
## Step:  AIC=3341.63
## medcond ~ diabetes + age + exercise + LDL
##
##            Df Deviance    AIC
## + drinkany  1   3325.8 3337.8
## <none>          3331.6 3341.6
```

```
## + BMI       1   3330.1 3342.1
## + HDL       1   3330.6 3342.6
## + weight    1   3331.3 3343.3
## + smoking   1   3331.3 3343.3
## + nonwhite  1   3331.6 3343.6
##
## Step:  AIC=3337.76
## medcond ~ diabetes + age + exercise + LDL + drinkany
##
##            Df Deviance    AIC
## <none>          3325.8 3337.8
## + HDL       1   3323.9 3337.9
## + BMI       1   3324.4 3338.4
## + weight    1   3325.5 3339.5
## + smoking   1   3325.5 3339.5
## + nonwhite  1   3325.6 3339.6
##
## Call:  glm(formula = medcond ~ diabetes + age + exercise + LDL + drinkany,
##     family = "binomial", data = HERS.all)
##
## Coefficients:
## (Intercept)     diabetes          age     exercise          LDL
##   -1.256566     0.449843     0.017738    -0.242926    -0.002859
##    drinkany
##   -0.209523
##
## Degrees of Freedom: 2570 Total (i.e. Null);  2565 Residual
## Null Deviance:     3384
## Residual Deviance: 3326  AIC: 3338
```

The last step gives the final model which is similar to the model created with drop in deviance test, except with AIC, no interactions were considered.

```
summary(glm(medcond ~ diabetes + age + exercise + LDL + drinkany, family = "binomial", data = HERS.all))

##                 Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -1.256566068 0.467463133 -2.688054 7.186981e-03
## diabetes     0.449843338 0.094256169  4.772561 1.818978e-06
## age          0.017737643 0.006308035  2.811913 4.924786e-03
## exercise    -0.242925702 0.085745347 -2.833107 4.609802e-03
## LDL         -0.002858851 0.001115801 -2.562152 1.040259e-02
## drinkany    -0.209522895 0.086667066 -2.417561 1.562493e-02
```

Automatic BIC

Note that the difference between AIC and BIC is the amount by which the loglikelihood gets penalized. BIC $= -2\ln$ likelihood $+ \ln(n)p$. Again, a small BIC is good because a big likelihood is good. However, the likelihood will continue to grow with more parameters, so we penalize the BIC by adding $\ln(n)p$ (where p is the number of parameters we estimate in the model and n is the number of observations.)

To adjust the AIC function to work for BIC, modify the constant on p. For AIC the constant is 2, for BIC the constant is $\ln(n)$.

```r
dim(HERS.all)
```

```
## [1] 2571   40
```

```r
# only the intercept
fmod1.bic = glm(medcond ~ 1, data=HERS.all, family="binomial")
# check the variables listed
step(fmod1.bic, ~ age + nonwhite + smoking + drinkany + exercise + weight +
       BMI + HDL + LDL + diabetes, data=HERS.all, direction="forward", k=log(2571))
```

```
## Start:  AIC=3391.63
## medcond ~ 1
##
##            Df Deviance    AIC
## + diabetes  1   3354.7 3370.4
## + drinkany  1   3372.1 3387.8
## + exercise  1   3373.7 3389.4
## <none>          3383.8 3391.6
## + age       1   3376.5 3392.2
## + LDL       1   3377.1 3392.8
## + BMI       1   3377.5 3393.2
## + weight    1   3381.4 3397.1
## + HDL       1   3383.6 3399.3
## + nonwhite  1   3383.7 3399.4
## + smoking   1   3383.7 3399.4
##
## Step:  AIC=3370.39
## medcond ~ diabetes
##
##            Df Deviance    AIC
## + age       1   3345.4 3368.9
## <none>          3354.7 3370.4
## + exercise  1   3347.3 3370.9
## + drinkany  1   3348.2 3371.8
## + LDL       1   3348.6 3372.2
## + HDL       1   3353.0 3376.6
## + BMI       1   3353.6 3377.1
## + nonwhite  1   3354.5 3378.1
## + smoking   1   3354.6 3378.2
## + weight    1   3354.7 3378.2
##
## Step:  AIC=3368.91
## medcond ~ diabetes + age
##
##            Df Deviance    AIC
## <none>          3345.4 3368.9
## + exercise  1   3337.7 3369.1
## + drinkany  1   3339.8 3371.2
## + LDL       1   3339.9 3371.4
## + BMI       1   3343.1 3374.5
## + HDL       1   3344.4 3375.8
## + smoking   1   3344.6 3376.0
```

```
## + weight    1   3344.7 3376.1
## + nonwhite  1   3345.3 3376.8
##
## Call:  glm(formula = medcond ~ diabetes + age, family = "binomial",
##     data = HERS.all)
##
## Coefficients:
## (Intercept)     diabetes          age
##    -1.95199      0.51705      0.01909
##
## Degrees of Freedom: 2570 Total (i.e. Null);  2568 Residual
## Null Deviance:     3384
## Residual Deviance: 3345  AIC: 3351
```

Because we are penalizing the likelihood much more now (because we have a large sample), the model does not allow for as many parameters. The final model is:

```
summary(glm(medcond ~ diabetes + age, family = "binomial", data = HERS.all))$coef
```

```
##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -1.95198743 0.424204833 -4.601521 4.194169e-06
## diabetes     0.51704967 0.092266041  5.603900 2.095816e-08
## age          0.01909063 0.006275772  3.041956 2.350461e-03
```

Backward Model Building

The drop1 function is equivalent to the add1 function. Additionally, the step function works with direction="forward" or direction="backward". One thing to note is that for drop1, you only need to provide the original model, and R will automatically know which variable to test.

Automatic Drop-in-deviance

```
HERS.all = HERS[complete.cases(HERS),]    # remove all rows with any NA values
bmod1 = glm(medcond ~ age + nonwhite + smoking + drinkany + exercise + weight +
       BMI + HDL + LDL + diabetes, data=HERS.all, family="binomial")  # only the intercept
drop1(bmod1, data=HERS.all, test="Chisq")    # check the variables listed

## Single term deletions
##
## Model:
## medcond ~ age + nonwhite + smoking + drinkany + exercise + weight +
##     BMI + HDL + LDL + diabetes
##          Df Deviance    AIC     LRT  Pr(>Chi)
## <none>        3319.6 3341.6
## age       1   3326.0 3346.0  6.4236   0.01126 *
## nonwhite  1   3319.9 3339.9  0.3519   0.55302
## smoking   1   3320.2 3340.2  0.5920   0.44165
## drinkany  1   3326.0 3346.0  6.4505   0.01109 *
## exercise  1   3325.7 3345.7  6.1216   0.01335 *
## weight    1   3321.2 3341.2  1.5762   0.20932
## BMI       1   3322.7 3342.7  3.1081   0.07790 .
## HDL       1   3322.0 3342.0  2.4275   0.11922
## LDL       1   3326.6 3346.6  7.0469   0.00794 **
## diabetes  1   3340.1 3360.1 20.5158 5.914e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

bmod2 = glm(medcond ~ age + smoking + drinkany + exercise + weight +
       BMI + HDL + LDL + diabetes, data=HERS.all, family="binomial")  # only the intercept
drop1(bmod2, data=HERS.all, test="Chisq")    # check the variables listed

## Single term deletions
##
## Model:
## medcond ~ age + smoking + drinkany + exercise + weight + BMI +
##     HDL + LDL + diabetes
##          Df Deviance    AIC     LRT  Pr(>Chi)
## <none>        3319.9 3339.9
## age       1   3326.9 3344.9  6.9222  0.008513 **
## smoking   1   3320.6 3338.6  0.6074  0.435784
## drinkany  1   3326.1 3344.1  6.1590  0.013074 *
## exercise  1   3326.0 3344.0  6.0515  0.013895 *
## weight    1   3321.4 3339.4  1.5010  0.220516
## BMI       1   3322.9 3340.9  2.9807  0.084265 .
## HDL       1   3322.2 3340.2  2.3037  0.129070
```

```
## LDL       1   3327.1 3345.1  7.2051  0.007270 **
## diabetes  1   3340.2 3358.2 20.2074 6.948e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

bmod3 = glm(medcond ~ age + drinkany + exercise + weight +
      BMI + HDL + LDL + diabetes, data=HERS.all, family="binomial")  # only the intercept
drop1(bmod3, data=HERS.all, test="Chisq")   # check the variables listed

## Single term deletions
##
## Model:
## medcond ~ age + drinkany + exercise + weight + BMI + HDL + LDL +
##     diabetes
##          Df Deviance    AIC     LRT Pr(>Chi)
## <none>        3320.6 3338.6
## age       1   3326.9 3342.9  6.3628 0.011654 *
## drinkany  1   3326.7 3342.7  6.1943 0.012816 *
## exercise  1   3327.4 3343.4  6.8895 0.008671 **
## weight    1   3322.1 3338.1  1.5025 0.220282
## BMI       1   3323.4 3339.4  2.8194 0.093133 .
## HDL       1   3322.8 3338.8  2.1997 0.138039
## LDL       1   3327.7 3343.7  7.1016 0.007701 **
## diabetes  1   3340.3 3356.3 19.7841 8.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

bmod4 = glm(medcond ~ age + drinkany + exercise +
      BMI + HDL + LDL + diabetes, data=HERS.all, family="binomial")  # only the intercept
drop1(bmod4, data=HERS.all, test="Chisq")   # check the variables listed

## Single term deletions
##
## Model:
## medcond ~ age + drinkany + exercise + BMI + HDL + LDL + diabetes
##          Df Deviance    AIC     LRT  Pr(>Chi)
## <none>        3322.1 3338.1
## age       1   3329.8 3343.8  7.7850  0.005268 **
## drinkany  1   3328.6 3342.6  6.5930  0.010238 *
## exercise  1   3329.0 3343.0  6.9679  0.008299 **
## BMI       1   3323.9 3337.9  1.8482  0.173996
## HDL       1   3324.4 3338.4  2.3500  0.125282
## LDL       1   3329.1 3343.1  7.0173  0.008073 **
## diabetes  1   3341.7 3355.7 19.5989 9.552e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

bmod5 = glm(medcond ~ age + drinkany + exercise +
      HDL + LDL + diabetes, data=HERS.all, family="binomial")  # only the intercept
drop1(bmod5, data=HERS.all, test="Chisq")   # check the variables listed

## Single term deletions
```

```
##
## Model:
## medcond ~ age + drinkany + exercise + HDL + LDL + diabetes
##          Df Deviance    AIC     LRT Pr(>Chi)
## <none>        3323.9 3337.9
## age       1   3330.9 3342.9  6.9660 0.008307 **
## drinkany  1   3330.6 3342.6  6.7206 0.009531 **
## exercise  1   3332.1 3344.1  8.2446 0.004087 **
## HDL       1   3325.8 3337.8  1.8567 0.173007
## LDL       1   3330.5 3342.5  6.6381 0.009982 **
## diabetes  1   3347.8 3359.8 23.8902 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
bmod6 = glm(medcond ~ age + drinkany + exercise +
      LDL + diabetes, data=HERS.all, family="binomial")  # only the intercept
drop1(bmod6, data=HERS.all, test="Chisq")   # check the variables listed
```

```
## Single term deletions
##
## Model:
## medcond ~ age + drinkany + exercise + LDL + diabetes
##          Df Deviance    AIC     LRT  Pr(>Chi)
## <none>        3325.8 3337.8
## age       1   3333.7 3343.7  7.9661  0.004766 **
## drinkany  1   3331.6 3341.6  5.8702  0.015400 *
## exercise  1   3333.8 3343.8  8.0780  0.004481 **
## LDL       1   3332.4 3342.4  6.6595  0.009863 **
## diabetes  1   3348.4 3358.4 22.6191 1.975e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The backward process resulted in the same model as the forward process:

```
summary(bmod6)$coef
```

```
##                 Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -1.256566068 0.467463133 -2.688054 7.186981e-03
## age          0.017737643 0.006308035  2.811913 4.924786e-03
## drinkany    -0.209522895 0.086667066 -2.417561 1.562493e-02
## exercise    -0.242925702 0.085745347 -2.833107 4.609802e-03
## LDL         -0.002858851 0.001115801 -2.562152 1.040259e-02
## diabetes     0.449843338 0.094256169  4.772561 1.818978e-06
```

Manual Backward Drop-in-Deviance

```
summary(glm(medcond ~ (age + diabetes + weight + drinkany)^2, family="binomial"))
```

```
##
## Call:
```

```
## glm(formula = medcond ~ (age + diabetes + weight + drinkany)^2,
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2716  -0.9756  -0.8573   1.3419   1.9508
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.1074581  2.1633263  -0.512   0.6087
## age                0.0085067  0.0316546   0.269   0.7881
## diabetes           1.8870002  1.1692772   1.614   0.1066
## weight            -0.0142558  0.0289592  -0.492   0.6225
## drinkany          -0.5868600  1.0756086  -0.546   0.5853
## age:diabetes      -0.0304417  0.0147880  -2.059   0.0395 *
## age:weight         0.0002082  0.0004289   0.486   0.6273
## age:drinkany       0.0073401  0.0131778   0.557   0.5775
## diabetes:weight    0.0078719  0.0062412   1.261   0.2072
## diabetes:drinkany -0.1361005  0.2052152  -0.663   0.5072
## weight:drinkany   -0.0016115  0.0061397  -0.262   0.7930
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3642.6  on 2758  degrees of freedom
## Residual deviance: 3585.7  on 2748  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 3607.7
##
## Number of Fisher Scoring iterations: 4
```

```
summary(glm(medcond ~ age + diabetes + weight + drinkany, family="binomial"))
```

```
##
## Call:
## glm(formula = medcond ~ age + diabetes + weight + drinkany, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2263  -0.9714  -0.8600   1.3403   1.7118
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.874672   0.504526  -3.716 0.000203 ***
## age          0.018362   0.006195   2.964 0.003038 **
## diabetes     0.432470   0.092422   4.679 2.88e-06 ***
## weight       0.001427   0.002854   0.500 0.617138
## drinkany    -0.252544   0.083452  -3.026 0.002476 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3642.6  on 2758  degrees of freedom
## Residual deviance: 3594.8  on 2754  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 3604.8
##
## Number of Fisher Scoring iterations: 4
```

- The big model (with all of the interaction terms) has a deviance of 3585.7; the additive model has a deviance of 3594.8.

$$\begin{aligned} \chi^2_6 &= 3594.8 - 3585.7 = 9.1 \\ p - value &= P(\chi^2_6 \geq 9.1) = 1 - pchisq(9.1, 6) = 0.1680318 \end{aligned}$$

We cannot reject the null hypothesis, so we know that we don't need the 6 interaction terms. Next we will check whether we need weight.

```
summary(glm(medcond ~ age + diabetes + drinkany, family="binomial"))
```

```
##
## Call:
## glm(formula = medcond ~ age + diabetes + drinkany, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2236  -0.9699  -0.8627   1.3443   1.7032
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.722505   0.412727  -4.173 3.00e-05 ***
## age          0.017573   0.006052   2.904  0.00369 **
## diabetes     0.442160   0.089529   4.939 7.86e-07 ***
## drinkany    -0.251505   0.083424  -3.015  0.00257 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3644.5  on 2760  degrees of freedom
## Residual deviance: 3597.3  on 2757  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 3605.3
##
## Number of Fisher Scoring iterations: 4
```

- The additive model has a deviance of 3594.8; the model without weight is 3597.3.

$$\begin{aligned} \chi^2_1 &= 3597.3 - 3594.8 = 2.5 \\ p - value &= P(\chi^2_1 \geq 2.5) = 1 - pchisq(2.5, 1) = 0.1138463 \end{aligned}$$

We cannot reject the null hypothesis, so we know that we don't need the weight in the model either.

Automatic Backward BIC

```
bmod1.aic = glm(medcond ~ age + nonwhite + smoking + drinkany + exercise + weight +
      BMI + HDL + LDL + diabetes, data=HERS.all, family="binomial")  # only the intercept
# check the variables listed
step(bmod1.aic, data=HERS.all, direction="backward", k=log(2571))

## Start:  AIC=3405.96
## medcond ~ age + nonwhite + smoking + drinkany + exercise + weight +
##     BMI + HDL + LDL + diabetes
##
##            Df Deviance    AIC
## - nonwhite  1   3319.9 3398.5
## - smoking   1   3320.2 3398.7
## - weight    1   3321.2 3399.7
## - HDL       1   3322.0 3400.5
## - BMI       1   3322.7 3401.2
## - exercise  1   3325.7 3404.2
## - age       1   3326.0 3404.5
## - drinkany  1   3326.0 3404.6
## - LDL       1   3326.6 3405.2
## <none>          3319.6 3406.0
## - diabetes  1   3340.1 3418.6
##
## Step:  AIC=3398.46
## medcond ~ age + smoking + drinkany + exercise + weight + BMI +
##     HDL + LDL + diabetes
##
##            Df Deviance    AIC
## - smoking   1   3320.6 3391.2
## - weight    1   3321.4 3392.1
## - HDL       1   3322.2 3392.9
## - BMI       1   3322.9 3393.6
## - exercise  1   3326.0 3396.7
## - drinkany  1   3326.1 3396.8
## - age       1   3326.9 3397.5
## - LDL       1   3327.1 3397.8
## <none>          3319.9 3398.5
## - diabetes  1   3340.2 3410.8
##
## Step:  AIC=3391.22
## medcond ~ age + drinkany + exercise + weight + BMI + HDL + LDL +
##     diabetes
##
##            Df Deviance    AIC
## - weight    1   3322.1 3384.9
## - HDL       1   3322.8 3385.6
## - BMI       1   3323.4 3386.2
## - drinkany  1   3326.7 3389.6
## - age       1   3326.9 3389.7
## - exercise  1   3327.4 3390.3
```

```
## - LDL       1   3327.7 3390.5
## <none>          3320.6 3391.2
## - diabetes  1   3340.3 3403.2
##
## Step:  AIC=3384.87
## medcond ~ age + drinkany + exercise + BMI + HDL + LDL + diabetes
##
##            Df Deviance    AIC
## - BMI       1   3323.9 3378.9
## - HDL       1   3324.4 3379.4
## - drinkany  1   3328.6 3383.6
## - exercise  1   3329.0 3384.0
## - LDL       1   3329.1 3384.0
## - age       1   3329.8 3384.8
## <none>          3322.1 3384.9
## - diabetes  1   3341.7 3396.6
##
## Step:  AIC=3378.87
## medcond ~ age + drinkany + exercise + HDL + LDL + diabetes
##
##            Df Deviance    AIC
## - HDL       1   3325.8 3372.9
## - LDL       1   3330.5 3377.7
## - drinkany  1   3330.6 3377.7
## - age       1   3330.9 3378.0
## <none>          3323.9 3378.9
## - exercise  1   3332.1 3379.3
## - diabetes  1   3347.8 3394.9
##
## Step:  AIC=3372.87
## medcond ~ age + drinkany + exercise + LDL + diabetes
##
##            Df Deviance    AIC
## - drinkany  1   3331.6 3370.9
## - LDL       1   3332.4 3371.7
## <none>          3325.8 3372.9
## - age       1   3333.7 3373.0
## - exercise  1   3333.8 3373.1
## - diabetes  1   3348.4 3387.6
##
## Step:  AIC=3370.89
## medcond ~ age + exercise + LDL + diabetes
##
##            Df Deviance    AIC
## - LDL       1   3337.7 3369.1
## <none>          3331.6 3370.9
## - exercise  1   3339.9 3371.4
## - age       1   3340.5 3371.9
## - diabetes  1   3359.2 3390.6
##
## Step:  AIC=3369.13
```

```
## medcond ~ age + exercise + diabetes
##
##            Df Deviance    AIC
## - exercise  1   3345.4 3368.9
## <none>          3337.7 3369.1
## - age       1   3347.3 3370.9
## - diabetes  1   3366.1 3389.7
##
## Step:  AIC=3368.91
## medcond ~ age + diabetes
##
##            Df Deviance    AIC
## <none>          3345.4 3368.9
## - age       1   3354.7 3370.4
## - diabetes  1   3376.5 3392.2
##
## Call:  glm(formula = medcond ~ age + diabetes, family = "binomial",
##     data = HERS.all)
##
## Coefficients:
## (Intercept)          age     diabetes
##    -1.95199      0.01909      0.51705
##
## Degrees of Freedom: 2570 Total (i.e. Null);  2568 Residual
## Null Deviance:     3384
## Residual Deviance: 3345  AIC: 3351
```

The last step gives the final model. Only age and diabetes is kept.

```
summary(glm(medcond ~ diabetes + age, family = "binomial", data = HERS.all))$coef
```

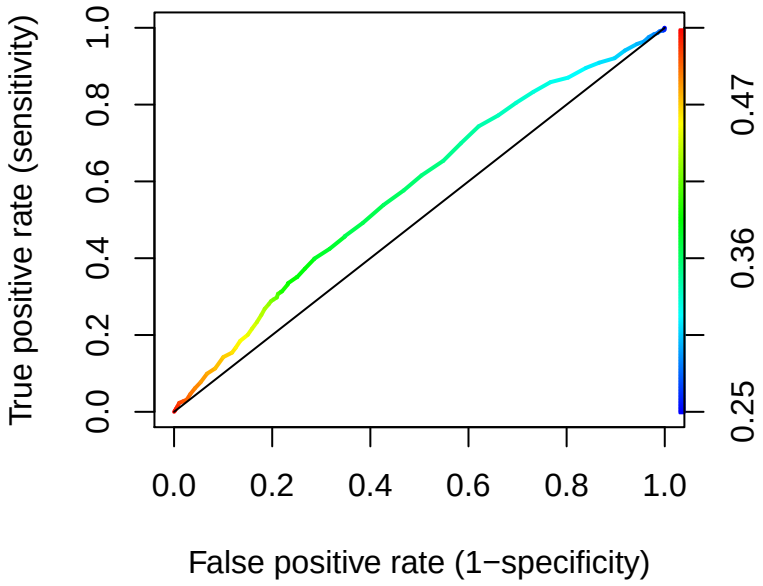
```
##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -1.95198743 0.424204833 -4.601521 4.194169e-06
## diabetes     0.51704967 0.092266041  5.603900 2.095816e-08
## age          0.01909063 0.006275772  3.041956 2.350461e-03
```

ROC Curves

Using the forward AIC model, compute and draw an ROC curve.

```
library(ROCR)
attach(HERS.all)

med.log = glm(medcond ~ diabetes + age, family = "binomial", data = HERS.all)
med.pred = prediction(fitted(med.log), medcond)
roc = performance(med.pred, "tpr", "fpr")
plot(roc, lwd=2, colorize=TRUE, xlab="False positive rate (1-specificity)",
     ylab="True positive rate (sensitivity)")
lines(x=c(0, 1), y=c(0, 1), col="black", lwd=1)
```




```
performance(med.pred, measure="auc")@y.values

## [[1]]
## [1] 0.5778021
```