BST 210 Project: Check-In 2

Daniel Herrera, Willow Duffell, Lauren Mock

11/6/2021

Group 4 Members:

Daniel Herrera
Willow Duffell
Lauren Mock

1) Logistic regression to predict survival

We want to predict the probability of survival after transplantation, given known covariates that can be measured before transplantation.

question: what determines dosage amount? so should we include this as a predictor or not?

```
# remove variables that are measured after transplantation
predictors <- bone[,c(1:24,37)] # selects only predictors and survival (the outcome)

# logistic model with all covariates
mod_all_vars <- glm(survival_status ~ ., family = binomial(), data = predictors)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(mod_all_vars)
```

```
##
## Call:
## glm(formula = survival_status ~ ., family = binomial(), data = predictors)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
##  -8.49     0.00     0.00     0.00     8.49
##
## Coefficients: (8 not defined because of singularities)
##                                  Estimate Std. Error    z value Pr(>|z|)
## (Intercept)                     5.225e+16  9.288e+08   56248972   <2e-16 ***
## donor_age                      -1.101e+13  1.259e+06   -8746490   <2e-16 ***
## donor_age_below_35yes          -5.032e+14  2.056e+07  -24476219   <2e-16 ***
## donor_ABOA                      8.323e+13  1.326e+07    6274183   <2e-16 ***
## donor_ABOAB                     4.244e+13  2.241e+07    1893738   <2e-16 ***
## donor_ABOB                     -1.613e+14  1.684e+07   -9577145   <2e-16 ***
```

```
## donor_CMVabsent                  -2.471e+15  6.966e+07  -35473036   <2e-16 ***
## donor_CMVpresent                 -2.322e+15  6.747e+07  -34408475   <2e-16 ***
## recipient_age                    -2.082e+14  3.746e+06  -55584889   <2e-16 ***
## recipient_age_below_10yes        -4.587e+15  7.250e+07  -63264479   <2e-16 ***
## recipient_age_int10_20           -3.893e+15  7.521e+07  -51763853   <2e-16 ***
## recipient_age_int5_10             7.178e+14  2.117e+07   33911823   <2e-16 ***
## recipient_gendermale             -4.457e+14  1.283e+07  -34736324   <2e-16 ***
## recipient_body_mass               3.943e+13  6.992e+05   56394815   <2e-16 ***
## recipient_ABO0                    5.992e+15  1.053e+08   56929168   <2e-16 ***
## recipient_ABOA                    6.374e+15  1.044e+08   61066102   <2e-16 ***
## recipient_ABOAB                   8.149e+15  1.069e+08   76217000   <2e-16 ***
## recipient_ABOB                    6.529e+15  1.056e+08   61844593   <2e-16 ***
## recipient_rhminus                -7.362e+15  7.767e+07  -94792174   <2e-16 ***
## recipient_rhplus                 -6.799e+15  7.607e+07  -89378726   <2e-16 ***
## recipient_CMVabsent              -3.751e+15  8.196e+07  -45766000   <2e-16 ***
## recipient_CMVpresent             -6.393e+15  6.178e+07 -103473019   <2e-16 ***
## diseaseAML                        2.363e+14  1.840e+07   12842621   <2e-16 ***
## diseasechronic                    5.964e+14  1.606e+07   37141089   <2e-16 ***
## diseaselymphoma                   3.793e+15  2.872e+07  132048777   <2e-16 ***
## diseasenonmalignant               3.783e+14  1.709e+07   22136973   <2e-16 ***
## disease_groupnonmalignant                NA         NA         NA       NA
## gender_matchother                -3.012e+14  1.658e+07  -18166742   <2e-16 ***
## ABO_matchmatched                  4.224e+14  1.357e+07   31117596   <2e-16 ***
## ABO_matchmismatched                      NA         NA         NA       NA
## CMV_status0                       3.104e+15  8.549e+07   36308361   <2e-16 ***
## CMV_status1                       2.987e+15  8.144e+07   36680277   <2e-16 ***
## CMV_status2                       6.102e+15  6.789e+07   89891007   <2e-16 ***
## CMV_status3                       5.696e+15  6.186e+07   92069476   <2e-16 ***
## HLA_match..out.of.10.            -4.401e+15  9.116e+07  -48277177   <2e-16 ***
## HLA_mismatchmismatched           -2.659e+15  8.186e+07  -32478517   <2e-16 ***
## antigen0                          1.592e+15  7.991e+07   19917980   <2e-16 ***
## antigen1                         -4.390e+15  2.072e+08  -21191309   <2e-16 ***
## antigen2                         -7.893e+15  2.735e+08  -28859725   <2e-16 ***
## antigen3                         -9.568e+15  3.512e+08  -27247229   <2e-16 ***
## allele0                                  NA         NA         NA       NA
## allele1                           4.897e+15  1.805e+08   27128269   <2e-16 ***
## allele2                           2.226e+15  1.025e+08   21723064   <2e-16 ***
## allele3                                  NA         NA         NA       NA
## allele4                                  NA         NA         NA       NA
## HLA_group_1matched                       NA         NA         NA       NA
## HLA_group_1mismatched                    NA         NA         NA       NA
## HLA_group_1one_allel             -1.152e+15  4.117e+07  -27983055   <2e-16 ***
## HLA_group_1one_antigen            3.266e+14  3.059e+07   10675726   <2e-16 ***
## HLA_group_1three_diffs            1.063e+15  4.363e+07   24364093   <2e-16 ***
## HLA_group_1two_diffs                     NA         NA         NA       NA
## risk_grouplow                    -3.750e+14  1.327e+07  -28262664   <2e-16 ***
## stem_cell_sourceperipheral_blood -1.213e+15  1.299e+07  -93362641   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:  254.51  on 184  degrees of freedom
## Residual deviance: 4397.33  on 140  degrees of freedom
```

```
##   (2 observations deleted due to missingness)
## AIC: 4487.3
##
## Number of Fisher Scoring iterations: 24
```

```
p_hats <- mod_all_vars$fitted.values

# only predicts 2 values, not sure why...

# regardless, we will definitely need to select some variables that make sense
```

Variables related to the donor

```
# variables related to the donor
mod_donor <- glm(survival_status ~ donor_age + donor_age_below_35 + donor_ABO + donor_CMV,
            family = binomial(), data = predictors)
summary(mod_donor)
```

```
##
## Call:
## glm(formula = survival_status ~ donor_age + donor_age_below_35 +
##     donor_ABO + donor_CMV, family = binomial(), data = predictors)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4147  -1.0962  -0.7793   1.1907   1.7256
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.11431    2.07739   1.018    0.309
## donor_age             -0.02419    0.03384  -0.715    0.475
## donor_age_below_35yes -0.79866    0.56299  -1.419    0.156
## donor_ABOA             0.24755    0.33994   0.728    0.466
## donor_ABOAB           -0.91205    0.70458  -1.294    0.196
## donor_ABOB            -0.09791    0.45209  -0.217    0.829
## donor_CMVabsent       -0.95773    1.57865  -0.607    0.544
## donor_CMVpresent      -1.25099    1.58615  -0.789    0.430
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.69  on 186  degrees of freedom
## Residual deviance: 250.38  on 179  degrees of freedom
## AIC: 266.38
##
## Number of Fisher Scoring iterations: 4
```

Variables related to the recipient

```
mod_recip <- glm(survival_status ~ recipient_age + recipient_gender + recipient_body_mass +
                   recipient_ABO + recipient_rh + recipient_CMV,
            family = binomial(), data = predictors)
summary(mod_recip)
```

```
##
## Call:
## glm(formula = survival_status ~ recipient_age + recipient_gender +
##     recipient_body_mass + recipient_ABO + recipient_rh + recipient_CMV,
##     family = binomial(), data = predictors)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7688  -0.9855  -0.7422   1.1409   1.8842
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             14.00011 1455.39766   0.010   0.9923
## recipient_age           -0.04094    0.07014  -0.584   0.5594
## recipient_gendermale     0.19051    0.32644   0.584   0.5595
## recipient_body_mass      0.03560    0.01949   1.826   0.0678 .
## recipient_ABOO           1.48006 2058.24309   0.001   0.9994
## recipient_ABOA           1.81733 2058.24305   0.001   0.9993
## recipient_ABOAB          1.83790 2058.24318   0.001   0.9993
## recipient_ABOB           1.72790 2058.24309   0.001   0.9993
## recipient_rhminus      -16.39755 1455.39764  -0.011   0.9910
## recipient_rhplus       -15.68622 1455.39758  -0.011   0.9914
## recipient_CMVabsent     -1.34419    0.68649  -1.958   0.0502 .
## recipient_CMVpresent    -1.14159    0.66511  -1.716   0.0861 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 254.51  on 184  degrees of freedom
## Residual deviance: 233.51  on 173  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 257.51
##
## Number of Fisher Scoring iterations: 14
```

```
mod_disease <- glm(survival_status ~ disease + disease_group,
            family = binomial(), data = predictors)
summary(mod_disease)
```

```
##
## Call:
## glm(formula = survival_status ~ disease + disease_group, family = binomial(),
##     data = predictors)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.1010  -1.0788  -0.9695   1.2793   1.4006
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.23639    0.24423  -0.968    0.333
## diseaseAML                  0.05407    0.42646   0.127    0.899
## diseasechronic             -0.07727    0.38826  -0.199    0.842
## diseaselymphoma            16.80246  799.84828   0.021    0.983
## diseasenonmalignant        -0.27444    0.43930  -0.625    0.532
## disease_groupnonmalignant        NA         NA      NA       NA
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.69  on 186  degrees of freedom
## Residual deviance: 242.43  on 182  degrees of freedom
## AIC: 252.43
##
## Number of Fisher Scoring iterations: 15
```

Variables related to the closeness of the match

```
mod_match <- glm(survival_status ~ gender_match + ABO_match + CMV_status + HLA_match..out.of.10. +
            antigen + allele,
            family = binomial(), data = predictors)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(mod_match)
```

```
##
## Call:
## glm(formula = survival_status ~ gender_match + ABO_match + CMV_status +
##     HLA_match..out.of.10. + antigen + allele, family = binomial(),
##     data = predictors)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -8.49   -8.49    0.00    0.00    8.49
##
## Coefficients: (3 not defined because of singularities)
##                          Estimate Std. Error   z value Pr(>|z|)
## (Intercept)           -1.169e+16  7.547e+08 -15485999   <2e-16 ***
## gender_matchother     -4.629e+14  1.375e+07 -33657953   <2e-16 ***
## ABO_matchmatched      -3.273e+15  7.148e+07 -45782797   <2e-16 ***
## ABO_matchmismatched   -4.546e+15  7.047e+07 -64513924   <2e-16 ***
## CMV_status0           -2.993e+12  2.047e+07   -146206   <2e-16 ***
## CMV_status1           -4.574e+14  2.190e+07 -20883191   <2e-16 ***
## CMV_status2            1.074e+14  1.990e+07   5395713   <2e-16 ***
## CMV_status3            4.129e+14  2.070e+07  19950018   <2e-16 ***
```

```
## HLA_match..out.of.10.  1.265e+15  7.492e+07  16885679   <2e-16 ***
## antigen0               3.541e+15  6.871e+07  51529432   <2e-16 ***
## antigen1               7.844e+15  1.783e+08  43994258   <2e-16 ***
## antigen2               7.998e+15  2.441e+08  32764239   <2e-16 ***
## antigen3               1.020e+16  3.173e+08  32144169   <2e-16 ***
## allele0                       NA         NA        NA       NA
## allele1               -2.548e+15  1.612e+08 -15802759   <2e-16 ***
## allele2               -1.831e+15  9.090e+07 -20147579   <2e-16 ***
## allele3                       NA         NA        NA       NA
## allele4                       NA         NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:  257.69  on 186  degrees of freedom
## Residual deviance: 6199.51  on 172  degrees of freedom
## AIC: 6229.5
##
## Number of Fisher Scoring iterations: 25
```

```
p_hats <- mod_match$fitted.values
head(p_hats, 30)
```

```
##            1            2            3            4            5            6
## 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16 1.000000e+00 2.220446e-16
##            7            8            9           10           11           12
## 2.220446e-16 2.220446e-16 2.220446e-16 1.000000e+00 1.000000e+00 1.000000e+00
##           13           14           15           16           17           18
## 2.220446e-16 2.220446e-16 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16
##           19           20           21           22           23           24
## 2.220446e-16 2.220446e-16 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##           25           26           27           28           29           30
## 1.000000e+00 2.220446e-16 1.000000e+00 1.000000e+00 2.220446e-16 1.000000e+00
```

```
mean(p_hats)
```

```
## [1] 0.6256684
```

```
# looks like something weird is happening with antigen (very bad predictor of survival)
mod_match <- glm(survival_status ~ gender_match + ABO_match + CMV_status + HLA_match..out.of.10. +
            allele,
           family = binomial(), data = predictors)
summary(mod_match)
```

```
##
## Call:
## glm(formula = survival_status ~ gender_match + ABO_match + CMV_status +
##     HLA_match..out.of.10. + allele, family = binomial(), data = predictors)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.5103  -1.0575  -0.8821   1.1982   1.5784
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.8028  2058.2496   0.000   0.9997
## gender_matchother        -0.0981     0.4159  -0.236   0.8135
## ABO_matchmatched        -14.7127  1455.3977  -0.010   0.9919
## ABO_matchmismatched     -15.2089  1455.3977  -0.010   0.9917
## CMV_status0              -0.7834     0.6240  -1.255   0.2093
## CMV_status1              -1.1655     0.6782  -1.719   0.0857 .
## CMV_status2              -0.7112     0.6064  -1.173   0.2408
## CMV_status3              -0.6753     0.6272  -1.077   0.2816
## HLA_match..out.of.10.     0.1709     0.5224   0.327   0.7435
## allele0                  14.6597  1455.3976   0.010   0.9920
## allele1                  15.1358  1455.3977   0.010   0.9917
## allele2                  14.9939  1455.3978   0.010   0.9918
## allele3                  15.5089  1455.3985   0.011   0.9915
## allele4                  31.2628  2058.2436   0.015   0.9879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.69  on 186  degrees of freedom
## Residual deviance: 248.14  on 173  degrees of freedom
## AIC: 276.14
##
## Number of Fisher Scoring iterations: 14
```

```
p_hats <- mod_match$fitted.values
head(p_hats, 30)
```

```
##         1         2         3         4         5         6         7         8
## 0.5199888 0.4929642 0.5110096 0.2877437 0.4454260 0.5644117 0.5644117 0.3612597
##         9        10        11        12        13        14        15        16
## 0.3888522 0.4524561 0.3223306 0.4910580 0.2877437 0.3718402 0.4929642 0.4722720
##        17        18        19        20        21        22        23        24
## 0.3223306 0.2877437 0.5644117 0.3974295 0.4967691 0.4282845 0.5199888 0.4722720
##        25        26        27        28        29        30
## 0.4929642 0.3950274 0.5174821 0.6446383 0.3888522 0.5199888
```

```
mean(p_hats)
```

```
## [1] 0.4545455
```

```
# when we remove antigen, it looks pretty normal
```

Variables related to stem cell source

```r
mod_source <- glm(survival_status ~ stem_cell_source,
            family = binomial(), data = predictors)
summary(mod_source)
```

```
##
## Call:
## glm(formula = survival_status ~ stem_cell_source, family = binomial(),
##     data = predictors)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.302  -1.045  -1.045   1.316   1.316
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       0.2877     0.3118   0.923   0.3562
## stem_cell_sourceperipheral_blood -0.6076     0.3543  -1.715   0.0863 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.69  on 186  degrees of freedom
## Residual deviance: 254.71  on 185  degrees of freedom
## AIC: 258.71
##
## Number of Fisher Scoring iterations: 4
```