Final Project Slides

Augustus Ge

April 21, 2019

Variable Explanation

- Our variables are log_mem_imm, dspan_for, dspan_back, animal, vegetable, trails_a, trails_b, dig_sym, bnt
- log_mem_imm
- dspan_for and dspan_back are forward and backward memorization of a sequence of digits
- animal and vegetable are the amount of unique animals and vegetables one can name in one minute
- trails_a and trails_b are connecting dots of digits and letters, total time to complete
- dig_sym is the digit symbol test substituting digits with their respective symbols, total time to complete
- bnt is the Boston Naming Test, subjects are to name drawings of common items, max of 30
- DEMENTED is the dementia status of subjects, with DEMENTED $= 1$ being dementia.

Modeling

- We create our first model using all of these variables

```
model1 = glm(DEMENTED ~ log_mem_imm + dspan_for + dspan_back +
               animal + vegetable + trails_a + trails_b + dig_sym +
              bnt, data = nacc_unique1, family= binomial)
summary_model1 = summary(model1)
summary_model1$coefficients
```

```
##                 Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept) -0.143449674 0.684757266  -0.2094898 8.340659e-01
## log_mem_imm -0.318352168 0.023219587 -13.7105009 8.784923e-43
## dspan_for    0.078851689 0.051876415   1.5199911 1.285132e-01
## dspan_back   0.099692535 0.051729832   1.9271769 5.395759e-02
## animal      -0.128671592 0.024554418  -5.2402623 1.603485e-07
## vegetable   -0.162863767 0.029249665  -5.5680557 2.575974e-08
## trails_a    -0.002203352 0.003498268  -0.6298408 5.287988e-01
## trails_b     0.007557559 0.001345184   5.6182356 1.929174e-08
## dig_sym     -0.020734436 0.005612127  -3.6945772 2.202530e-04
## bnt          0.057049536 0.018241432   3.1274703 1.763177e-03
```

```
cat("AIC: ", summary_model1$aic)
```

```
## AIC:  1047.118
```

- From our summary, we see that log_mem_imm has the most significant p-value. Additionally, dspan_for, dspan_back, and trails_a are not significant at the .05 level.
- AIC is 1047.1

Likelihood Ratio Test

- We compare two models, one with all of the test components and one with all of the test components minus the log_mem_imm.
- With a chisq test statistic of 246.79, we reject the null hypothesis and conclude that the model including the log_mem_imm test is the better statistical model.

```
model2 = glm(DEMENTED ~ dspan_for + dspan_back + animal + vegetable
        + trails_a + trails_b + dig_sym + bnt, data = nacc_unique1, family= binomial)
lrtest(model1, model2)
```

```
## Likelihood ratio test
##
## Model 1: DEMENTED ~ log_mem_imm + dspan_for + dspan_back + animal + vegetable +
##     trails_a + trails_b + dig_sym + bnt
## Model 2: DEMENTED ~ dspan_for + dspan_back + animal + vegetable + trails_a +
##     trails_b + dig_sym + bnt
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  10 -513.56
## 2   9 -636.96 -1 246.79  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Stepwise Test

- Running a step test on our model shows us how our AIC would be affected if we were to drop different variables. If we were to drop log_mem_imm, our AIC would go up by more than 200 points.
- Also, dropping trails_a actually decreases our AIC. The rest of the variables are helping our AIC.

```
step(model1, trace = 1)
```

```
## Start:  AIC=1047.12
## DEMENTED ~ log_mem_imm + dspan_for + dspan_back + animal + vegetable +
##     trails_a + trails_b + dig_sym + bnt
##
##               Df Deviance    AIC
## - trails_a     1   1027.5 1045.5
## <none>             1027.1 1047.1
## - dspan_for    1   1029.4 1047.4
## - dspan_back   1   1030.8 1048.8
## - bnt          1   1037.2 1055.2
## - dig_sym      1   1040.0 1058.0
## - animal       1   1056.3 1074.3
## - trails_b     1   1058.2 1076.2
## - vegetable    1   1060.2 1078.2
## - log_mem_imm  1   1273.9 1291.9
##
## Step:  AIC=1045.51
## DEMENTED ~ log_mem_imm + dspan_for + dspan_back + animal + vegetable +
##     trails_b + dig_sym + bnt
##
```

Modeling with only log_mem_imm

- With just one variable we can get our AIC to 1297.9, wheras using all of our variables minus log_mem_imm will give us an AIC of 1291.9

```
model3 = glm(DEMENTED ~ log_mem_imm, data = nacc_unique1, family = binomial)
summary(model3)
```

```
##
## Call:
## glm(formula = DEMENTED ~ log_mem_imm, family = binomial, data = nacc_unique1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3097  -0.1725  -0.0703  -0.0358   3.9509
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.30571    0.12122   2.522   0.0117 *
## log_mem_imm -0.45057    0.01992 -22.616   <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: 2197.8  on 7684  degrees of freedom
## Residual deviance: 1293.9  on 7683  degrees of freedom
## AIC: 1297.9
##
## Number of Fisher Scoring iterations: 8
```

Predicting with log_mem_imm Model

Finally, using the formula

$$y = 3.057 + -.451 * log\_mem\_imm$$

taking the inverse logit of y, and sending those with a probability greater than .5 to having a dementia status of 1 and the rest to 0, we can achieve an accuracy of .9733. However, we must consider that 96.76% of the subjects do not have dementia.