

Diagnosis of Dementia and History of Traumatic Brain Injury (TBI)

Bowei Zhang

2023-03-30

Data Wrangling

Data Cleaning

```
# import data
donor_raw <- read_csv("../data/DonorInformation.csv")

## Rows: 107 Columns: 19
## -- Column specification -----
## Delimiter: ","
## chr (11): name, age, sex, apo_e4_allele, longest_loc_duration, dsm_iv_clinic...
## dbl (8): donor_id, education_years, age_at_first_tbi, cerad, num_tbi_w_loc,...
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.

donor <- donor_raw %>%
  dplyr::select(-donor_id, -name, - race, - hispanic)

# Recode
# recode outcome dsm_iv_clinical_diagnosis
donor <- donor %>% mutate(dementia = case_when(
  dsm_iv_clinical_diagnosis == 'No Dementia' ~ 0,
  TRUE ~ 1,
)) %>% dplyr::select(-dsm_iv_clinical_diagnosis, - cerad, -control_set, -braak, - nia_reagan)

# recode 'age':
donor <- donor %>%
  mutate(age = case_when(
    age == '100+' ~ '100',
    age == '95-99' ~ '97',
    age == '90-94' ~ '92',
    TRUE ~ age
  )
) %>%
  mutate(age = as.numeric(age))
```

```

## Relabel column 'sex'
donor$sex <- ifelse(donor$sex == 'M', 1, 0)
donor$sex <- factor(donor$sex,
                    levels = 0:1,
                    labels = c('Female', 'Male'))

## relabel column 'apo_e4_allele'
donor = donor %>% mutate(apo_e4_allele = case_when(
  apo_e4_allele == 'N' ~ 0,
  apo_e4_allele == 'Y' ~ 1,
  apo_e4_allele == 'N/A' ~ 2
))
donor$apo_e4_allele <- factor(donor$apo_e4_allele,
                             levels = 0:2,
                             labels = c('not carrier', 'carrier', 'unknown'))

## relabel column 'longest_loc_duration'
donor = donor %>% mutate(longest_loc_duration = case_when(
  longest_loc_duration == "Unknown or N/A" ~ 0,
  longest_loc_duration == "< 10 sec" ~ 1,
  longest_loc_duration == "10 sec - 1 min" ~ 2,
  longest_loc_duration == "1-2 min" ~ 3,
  longest_loc_duration == "3-5 min" ~ 4,
  longest_loc_duration == "6-9 min" ~ 5,
  longest_loc_duration == "10 min - 1 hr" ~ 6,
  longest_loc_duration == "> 1 hr" ~ 7
))
donor$longest_loc_duration <- factor(donor$longest_loc_duration,
                                    levels = 0:7,
                                    labels = c(
  "Unknown or N/A",
  "< 10 sec",
  "10 sec - 1 min",
  "1-2 min",
  "3-5 min",
  "6-9 min",
  "10 min - 1 hr",
  "> 1 hr"
))

```

Explanatory Analysis

Backward Elimination on confounders: sex, apo_e4, education years, using $p = 0.2$

```

m.confounders <- logistf(data = donor,
                        formula = dementia ~ sex + apo_e4_allele + education_years,
                        family = binomial)
backward(m.confounders, slstay = 0.2)

```

```
## Step 0 : starting model
```

```
## Step 1 : removed sex (P= 0.7708075 )

## logistf(formula = dementia ~ apo_e4_allele + education_years,
## data = donor, family = binomial)
## Model fitted by Penalized ML
## Confidence intervals and p-values by Profile Likelihood
##
## Coefficients:
## (Intercept) apo_e4_allelecarrier apo_e4_alleleunknown
## 1.2120002 1.0056298 1.2077485
## education_years
## -0.1118727
##
## Likelihood ratio test=8.342034 on 3 df, p=0.0394473, n=107
```

```
## apoe4 and education yr
m1 <- glm(dementia ~ apo_e4_allele + education_years,
data = donor %>% filter(apo_e4_allele != 'unknown'),
family = binomial)
summary(m1)
```

```
##
## Call:
## glm(formula = dementia ~ apo_e4_allele + education_years, family = binomial,
## data = donor %>% filter(apo_e4_allele != "unknown"))
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -1.5980 -1.0265 -0.7073 1.2249 1.7368
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.41852 0.95295 1.489 0.1366
## apo_e4_allelecarrier 1.06086 0.53097 1.998 0.0457 *
## education_years -0.12746 0.06711 -1.899 0.0575 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 137.99 on 99 degrees of freedom
## Residual deviance: 130.50 on 97 degrees of freedom
## AIC: 136.5
##
## Number of Fisher Scoring iterations: 4
```

Investigate main effects

Age at first TBI

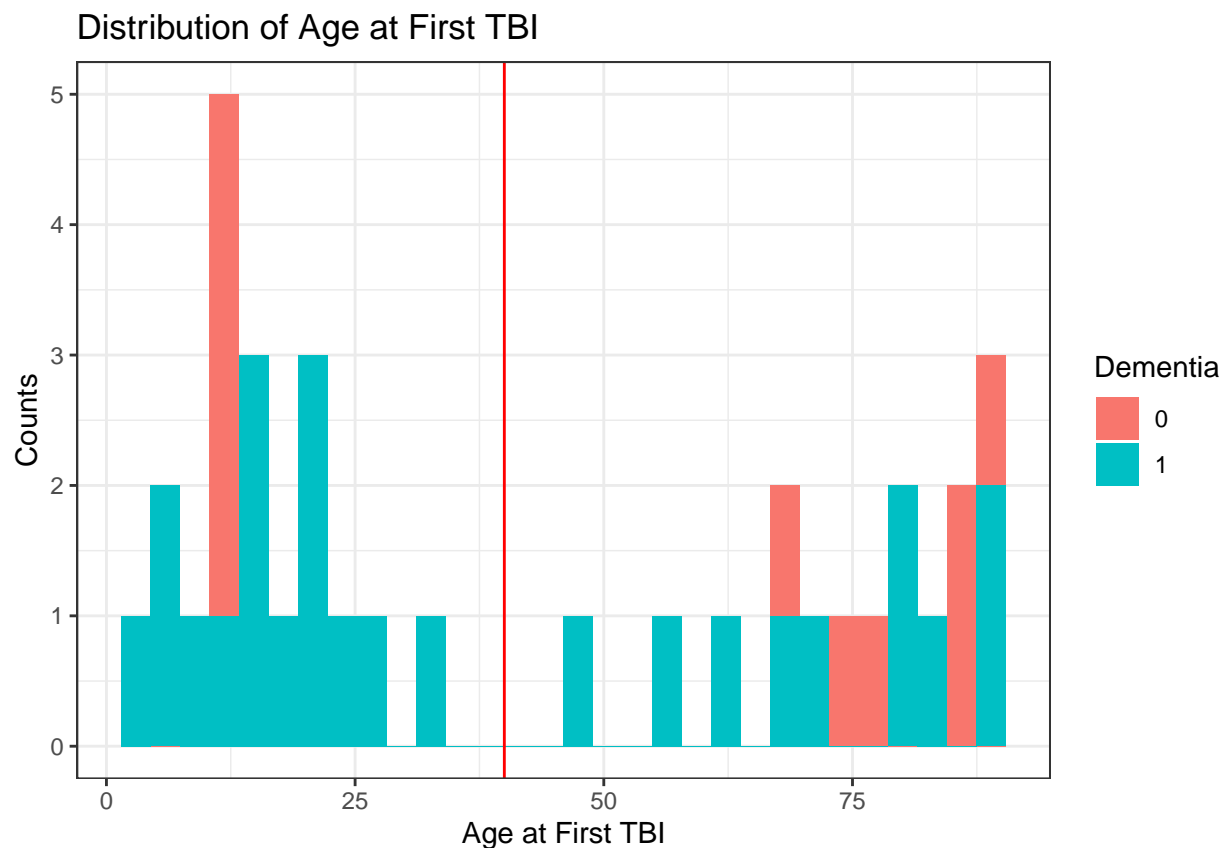
```

# create a donor 2 where apo_e4 and age_at_first_tbi are not missing
donor2 <- donor %>% filter(apo_e4_allele != 'unknown', age_at_first_tbi != 0)

# See the distribution of Age at First TBI
ggplot(data = donor2,
       aes(age_at_first_tbi, fill = as.factor(dementia)))+
  geom_histogram(position = "identity")+
  scale_fill_discrete(name = "Dementia") +
  geom_vline(xintercept = 40, col = "red") +
  theme_bw() +
  labs(x = "Age at First TBI",
       y = "Counts",
       title = "Distribution of Age at First TBI")

```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



```

# split into 3 age group: never, < 40 and >= 40
donor2 <- donor %>% mutate(age_at_first_tbi = case_when(
  age_at_first_tbi == 0 ~ 0,
  age_at_first_tbi < 40 ~ 1,
  TRUE ~ 2
)) %>%
  filter(apo_e4_allele != 'unknown')

```

```
donor2$age_at_first_tbi <- factor(donor2$age_at_first_tbi,
                                  levels = 0:2,
                                  labels = c("never", "before 40", "after 40"))

m2 <- glm(dementia ~ apo_e4_allele + education_years + age_at_first_tbi,
          data = donor2,
          family = binomial)

summary(m2)
```

```
##
## Call:
## glm(formula = dementia ~ apo_e4_allele + education_years + age_at_first_tbi,
##      family = binomial, data = donor2)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7777  -1.0356  -0.7691   1.1288   1.7725
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.25524    0.98724   1.271  0.2036
## apo_e4_allelecarrier  1.09816    0.53739   2.043  0.0410 *
## education_years    -0.12899    0.06745  -1.913  0.0558 .
## age_at_first_tbibefore 40  0.54408    0.49578   1.097  0.2725
## age_at_first_tbiafter 40  0.11591    0.54172   0.214  0.8306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 137.99  on 99  degrees of freedom
## Residual deviance: 129.26  on 95  degrees of freedom
## AIC: 139.26
##
## Number of Fisher Scoring iterations: 4
```

Longest Loss of Consciousness Duration

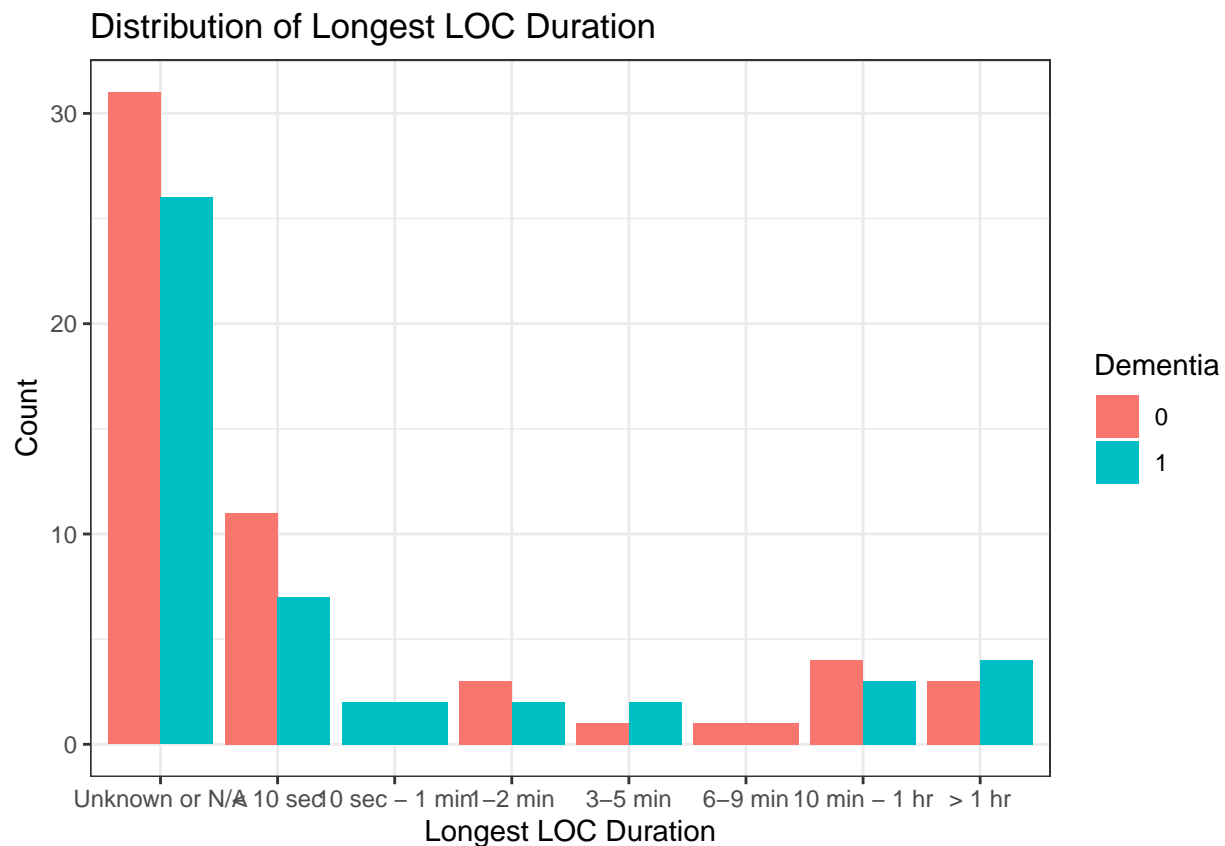
```
## Quasi-complete separation problem
donor2 %>%
  filter(ever_tbi_w_loc == 'N') %>%
  dplyr::select(ever_tbi_w_loc,
                age_at_first_tbi,
                longest_loc_duration,
                num_tbi_w_loc) %>%
  head(n=5)
```

```
## # A tibble: 5 x 4
##   ever_tbi_w_loc age_at_first_tbi longest_loc_duration num_tbi_w_loc
##   <chr>          <fct>          <fct>          <dbl>
## 1 N            never            Unknown or N/A            0
```

```
## 2 N          never          Unknown or N/A          0
## 3 N          never          Unknown or N/A          0
## 4 N          never          Unknown or N/A          0
## 5 N          never          Unknown or N/A          0
```

```
## longest_loc_duration
ggplot(data = donor2,
       aes(longest_loc_duration, fill = as.factor(dementia)))+
  geom_bar(position = "dodge", binwidth = 0.5)+
  scale_fill_discrete(name = "Dementia") +
  theme_bw() +
  labs(x = "Longest LOC Duration",
       y = "Count",
       title = "Distribution of Longest LOC Duration")
```

```
## Warning in geom_bar(position = "dodge", binwidth = 0.5): Ignoring unknown
## parameters: 'binwidth'
```



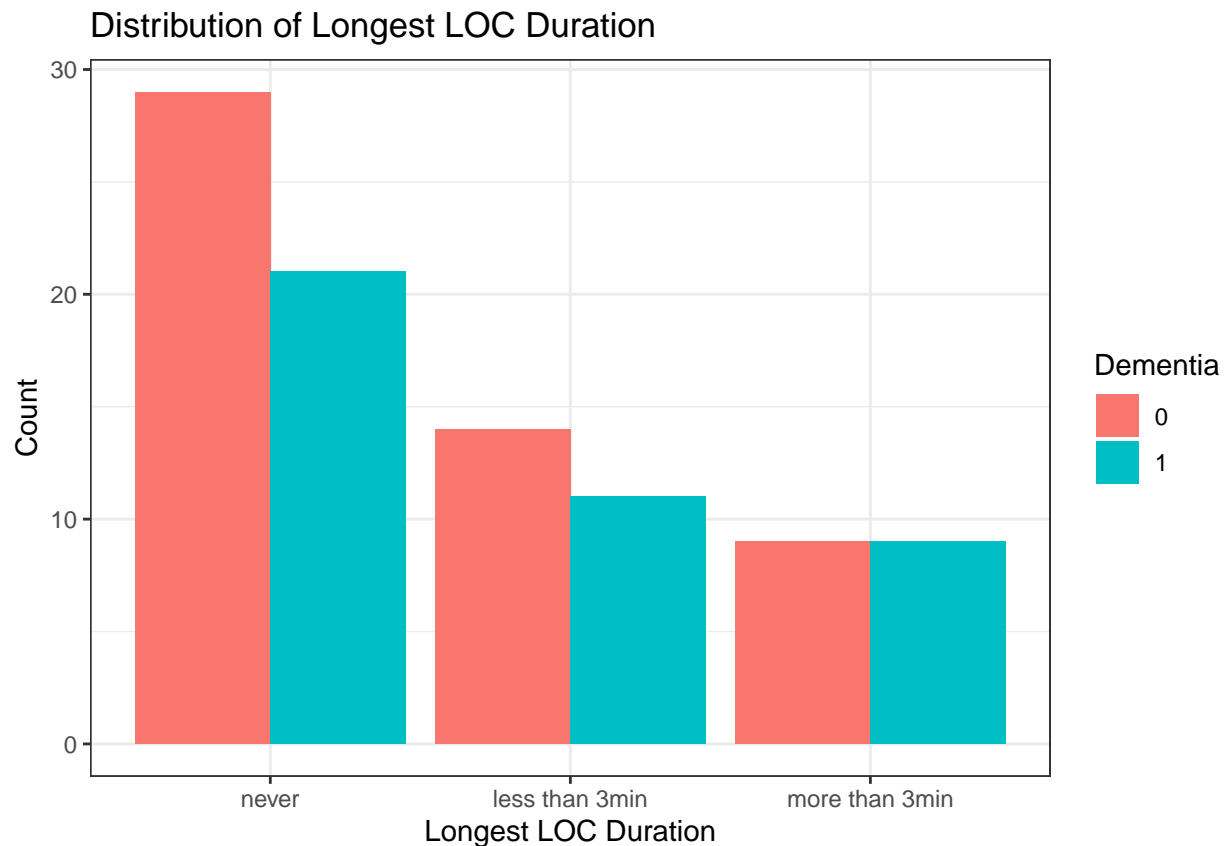
```
# recode longest_loc_duration
donor3 <- donor2 %>% mutate(longest_loc_duration = case_when(
  num_tbi_w_loc == 0 & longest_loc_duration == "Unknown or N/A" ~ 0,
  num_tbi_w_loc != 0 & longest_loc_duration == "Unknown or N/A" ~ 100, # missing values
  longest_loc_duration == "< 10 sec" ~ 1,
  longest_loc_duration == "10 sec - 1 min" ~ 1,
  longest_loc_duration == "1-2 min" ~ 1,
```

```

longest_loc_duration == "3-5 min" ~ 2,
longest_loc_duration == "6-9 min" ~ 2,
longest_loc_duration == "10 min - 1 hr" ~ 2,
longest_loc_duration == "> 1 hr" ~ 2
)) %>%
  filter(longest_loc_duration != 100)
donor3$longest_loc_duration <- factor(donor3$longest_loc_duration,
                                     levels = 0:2,
                                     labels = c("never", "less than 3min", "more than 3min"))

# See the new distribution
ggplot(data = donor3,
       aes(longest_loc_duration, fill = as.factor(dementia)))+
  geom_bar(position = "dodge")+
  scale_fill_discrete(name = "Dementia")+
  theme_bw() +
  labs(x = "Longest LOC Duration",
       y = "Count",
       title = "Distribution of Longest LOC Duration")

```



```

m3 <- glm(dementia ~ apo_e4_allele + education_years + longest_loc_duration,
          data = donor3,
          family = binomial)
summary(m3)

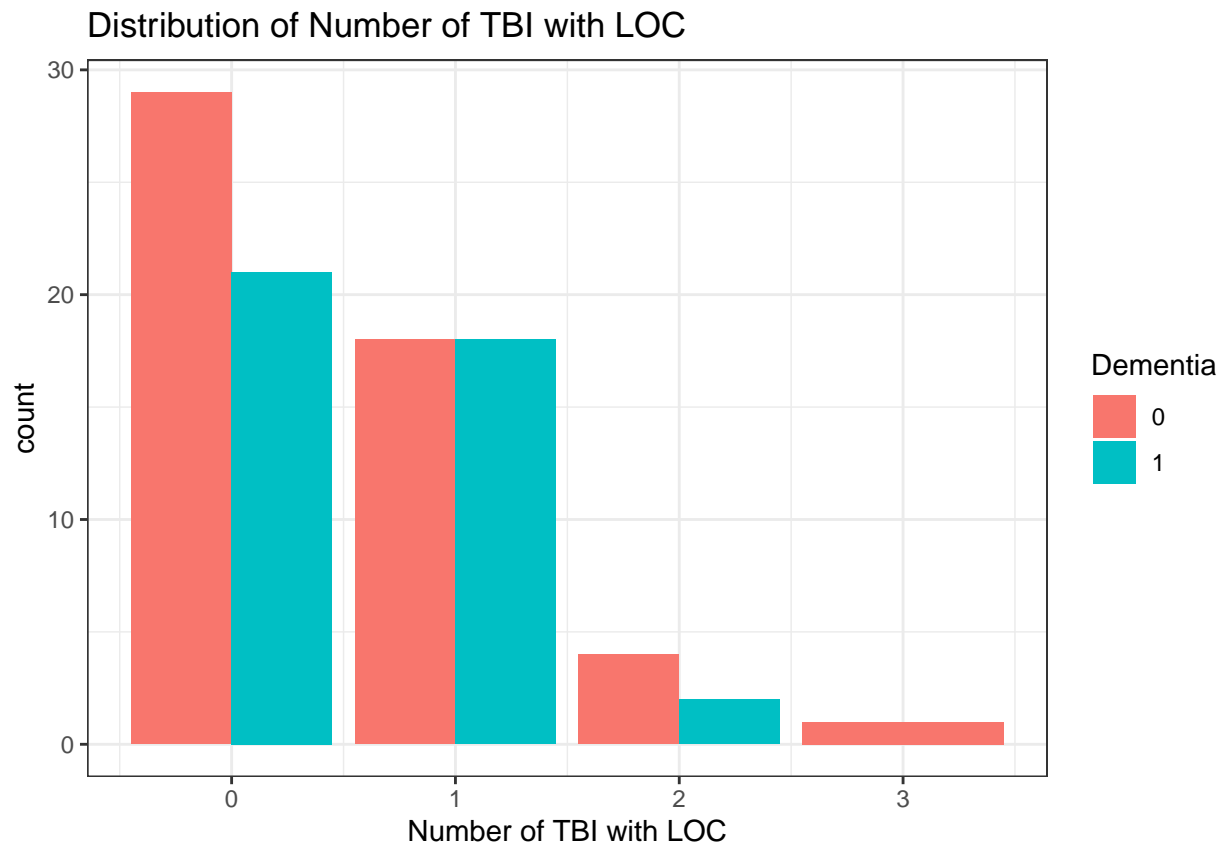
```

##

```
## Call:
## glm(formula = dementia ~ apo_e4_allele + education_years + longest_loc_duration,
##      family = binomial, data = donor3)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7025  -0.9985  -0.7418   1.1312   1.8490
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.39332    1.03091   1.352   0.1765
## apo_e4_allelecarrier    1.33291    0.57556   2.316   0.0206 *
## education_years       -0.14263    0.07111  -2.006   0.0449 *
## longest_loc_durationless than 3min  0.09227    0.52434   0.176   0.8603
## longest_loc_durationmore than 3min  0.45230    0.58543   0.773   0.4398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 127.62  on 92  degrees of freedom
## Residual deviance: 117.98  on 88  degrees of freedom
## AIC: 127.98
##
## Number of Fisher Scoring iterations: 4
```

Number of TBI with LOC

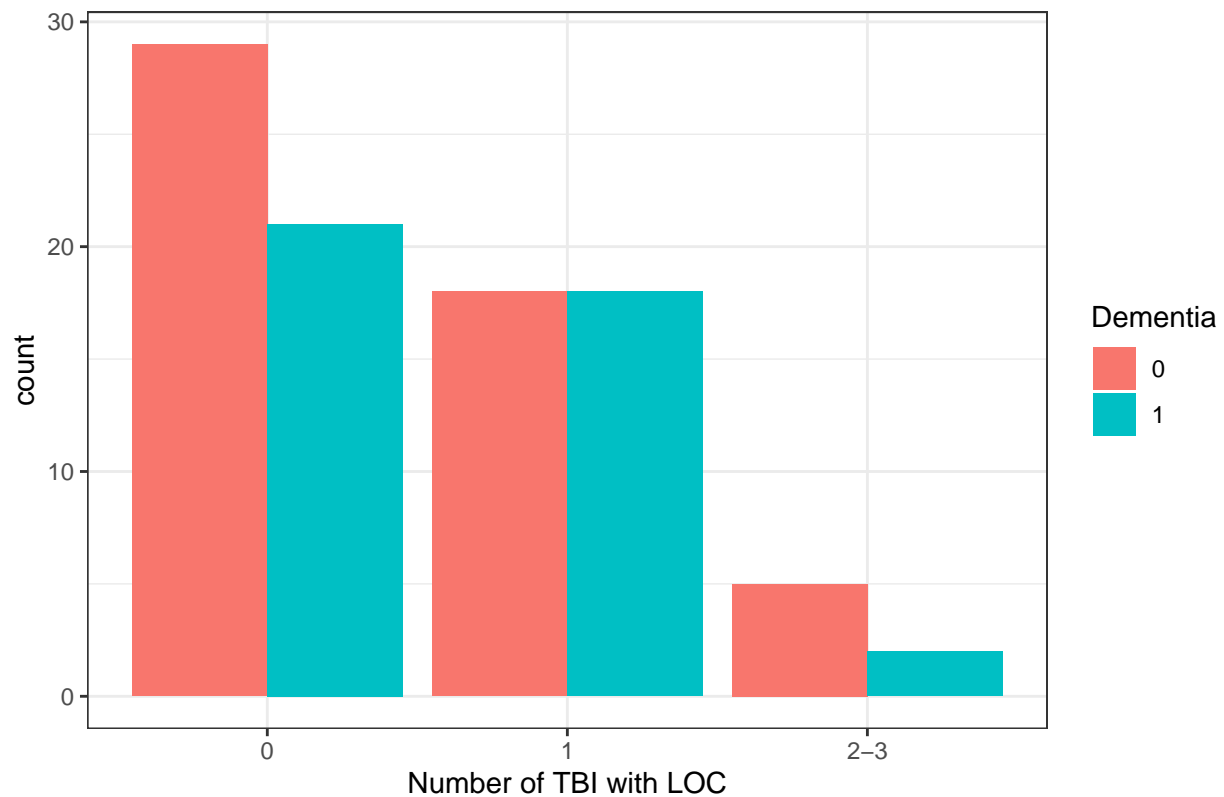
```
# Distribution of num_tbi_w_loc
ggplot(data = donor3,
       aes(num_tbi_w_loc, fill = as.factor(dementia)))+
  geom_bar(position = "dodge")+
  scale_fill_discrete(name = "Dementia") +
  theme_bw() +
  labs(x = "Number of TBI with LOC",
       title = "Distribution of Number of TBI with LOC")
```

```
# maybe we can combine group 2 and 3
donor4 <- donor3 %>% mutate(num_tbi_w_loc = case_when(
  num_tbi_w_loc >= 2 ~ 2,
  num_tbi_w_loc == 1 ~ 1,
  num_tbi_w_loc == 0 ~ 0,
))
donor4$num_tbi_w_loc <- factor(donor4$num_tbi_w_loc,
                             levels = 0:2,
                             labels = c('0', '1', '2-3'))

# plot
#donor4 %>% count(dementia, num_tbi_w_loc)
ggplot(data = donor4,
       aes(num_tbi_w_loc, fill = as.factor(dementia)))+
  geom_bar(position = "dodge")+
  scale_fill_discrete(name = "Dementia") +
  theme_bw() +
  labs(x = "Number of TBI with LOC",
       title = "Distribution of Number of TBI with LOC")
```

Distribution of Number of TBI with LOC



```
m5 <- glm(dementia ~ apo_e4_allele + education_years +
           num_tbi_w_loc,
           data = donor4,
           family = binomial)
summary(m5)
```

```
##
## Call:
## glm(formula = dementia ~ apo_e4_allele + education_years + num_tbi_w_loc,
##      family = binomial, data = donor4)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.740  -0.950  -0.745   1.137   1.732
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.32926    1.04040   1.278  0.2014
## apo_e4_allelecarrier 1.27803    0.57923   2.206  0.0274 *
## education_years   -0.13718    0.07198  -1.906  0.0567 .
## num_tbi_w_loc1     0.30413    0.46474   0.654  0.5128
## num_tbi_w_loc2-3   -0.09344    0.91362  -0.102  0.9185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 127.62 on 92 degrees of freedom
## Residual deviance: 118.09 on 88 degrees of freedom
## AIC: 128.09
##
## Number of Fisher Scoring iterations: 4
```

Model Fitting

Interaction terms

```
m2.apoe4int <- update(m2, . ~ . + apo_e4_allele:age_at_first_tbi)
anova(m2, m2.apoe4int, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: dementia ~ apo_e4_allele + education_years + age_at_first_tbi
## Model 2: dementia ~ apo_e4_allele + education_years + age_at_first_tbi +
## apo_e4_allele:age_at_first_tbi
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1 95 129.26
## 2 93 126.45 2 2.8145 0.2448
```

```
m2.eduyint <- update(m2, . ~ . + education_years:age_at_first_tbi)
anova(m2, m2.eduyint, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: dementia ~ apo_e4_allele + education_years + age_at_first_tbi
## Model 2: dementia ~ apo_e4_allele + education_years + age_at_first_tbi +
## education_years:age_at_first_tbi
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1 95 129.26
## 2 93 128.59 2 0.66942 0.7155
```

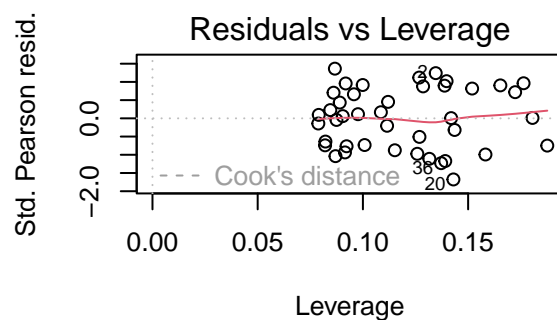
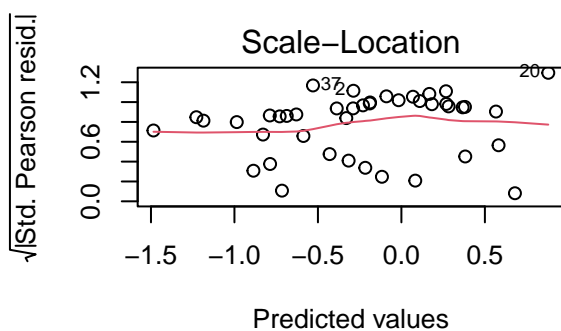
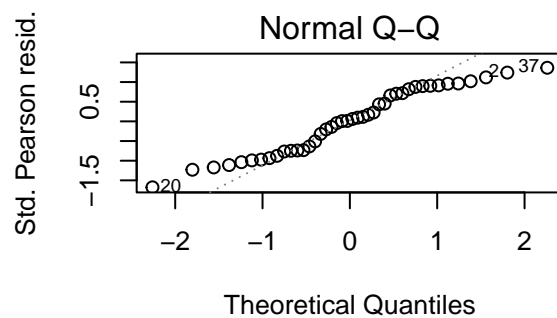
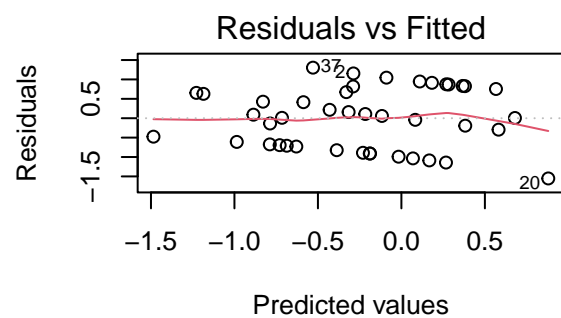
```
# aggregate model
m.final <- m2
ag.donor2 <- aggregate(dementia ~ apo_e4_allele + education_years + age_at_first_tbi,
                      data = donor2,
                      FUN = sum)
ag.donor2 <- cbind(ag.donor2,
                  aggregate(dementia ~ apo_e4_allele + education_years + age_at_first_tbi,
                            data = donor2,
                            FUN = length))
colnames(ag.donor2)[8] = 'tot'
ag.m.final = glm(dementia/tot ~ apo_e4_allele + education_years + age_at_first_tbi,
                 data = ag.donor2,
                 family = binomial)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

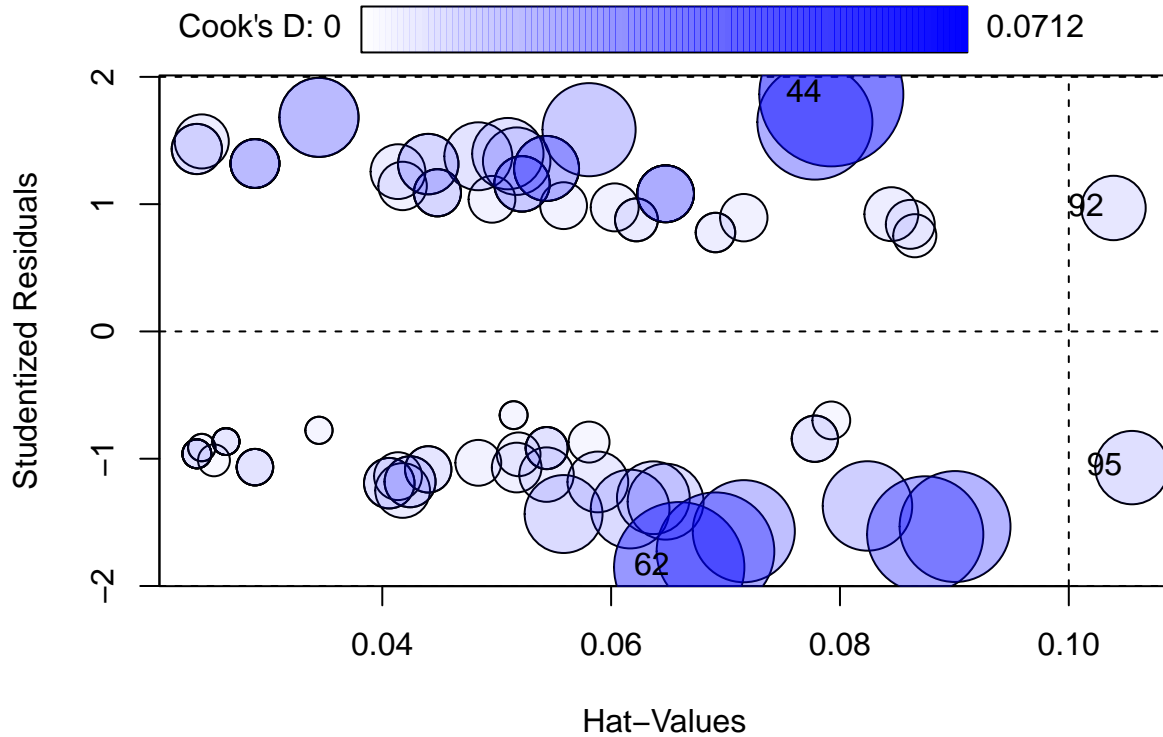
```
summary(ag.m.final)
```

```
##
## Call:
## glm(formula = dementia/tot ~ apo_e4_allele + education_years +
##      age_at_first_tbi, family = binomial, data = ag.donor2)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56638  -0.89900   0.03434   0.79535   1.40896
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.61066    1.40619   0.434   0.664
## apo_e4_allelecarrier  0.69687    0.71634   0.973   0.331
## education_years    -0.09979    0.09336  -1.069   0.285
## age_at_first_tbibefore 40  0.76971    0.79547   0.968   0.333
## age_at_first_tbiafter 40   0.25688    0.78688   0.326   0.744
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 35.523  on 41  degrees of freedom
## Residual deviance: 32.685  on 37  degrees of freedom
## AIC: 60.486
##
## Number of Fisher Scoring iterations: 3
```

```
# Diagnostic Plot
par(mfrow = c(2,2))
plot(ag.m.final)
```



```
# checking for influential points
influencePlot(m.final)
```



##	StudRes	Hat	CookD
## 44	1.8626613	0.07924397	0.07123272
## 62	-1.8527045	0.06595006	0.05829096
## 92	0.9699451	0.10390859	0.01424374
## 95	-1.0700025	0.10552021	0.01848570

Results

For the purpose of this analysis, it was our interest to find an association between an earlier exposure(s) to TBI and a later development of dementia. After conducting a full analysis of the Aging, Dementia and Traumatic Brain Injury (TBI) Project, originally collected from the ACT study, we fail to reject our null hypothesis. Thus, we conclude that there is not enough statistically significant evidence that suggests there is an association between TBI exposure and the diagnosis of dementia. Though it was not the question of interest, an association between carriers of the APOE4 gene and education years was found to be significant in the development of dementia. Below, Figure 5 shows us the distribution of diagnosis of dementia, education years, and APOE4 gene status in light of age at first TBI. The plot demonstrates that the odds of developing dementia decrease by 12% (95% CI: -23.5%, -0.08%) for every additional year of education, when all other factors are held constant. The odds of developing dementia increase by a factor of 2.99 (95% CI: 1.076, 9.067) for those who are carriers of the APOE4 gene, compared to those who do not carrier the gene, when all other factors are held constant. Figure 5 also illustrates that the odds of developing dementia for those who experienced their first TBI before the age of 40 is increased by 72.3% (95% CI: 0.655, 4.631) compared to those who had never experienced a TBI, when all other factors all held constant. Similarly, the odds of developing dementia for those who experienced their first TBI after the age of 40 is increased by 12.3% (95% CI: 0.383, 3.256) compared to those who had never experienced a TBI, when all other factors all held

constant. However, neither exposure to TBI before or after 40 years of age have significant results. Thus, we suggest further testing with a larger sample in order to see if an association between TBI exposure and dementia exists.

```
pred.m2 <- data.frame(expand.grid(education_years = c(6:21),
                                  apo_e4_allele = c("carrier", "not carrier"),
                                  age_at_first_tbi = c("never", "before 40", "after 40")))

ggplot(data = pred.m2,
       aes(x = education_years, y = predict(m2, newdata = pred.m2, type = "response"), col = apo_e4_allele)) +
  facet_wrap(facets = vars(age_at_first_tbi)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  theme_bw() +
  labs(x = "Education Years",
       y = "Predicted Rate of Dementia",
       title = "The Effect of Age at First TBI on Diagnosis of Dementia")
```

```
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
```

