Determining the Strongest Early Risk Factor Interaction for Early Onset Dementia

DS 400, Faith Hardie, Elora Tonaki

What are the Strongest Variable Interactions and Predictors for Early-Onset Dementia?

Based on our research question, we chose to focus on these variables:

- Age
- Sex
- MMSE (Mini Mental State Evaluation)
- CDR (Clinical Dementia Rating)
- nWBV (Normalized Whole Brain Volume)

We hypothesize that lower MMSE and nWBV strongly increase the probability of dementia, and that the predictive effect of reduced brain volume is even stronger at younger ages.

Methods

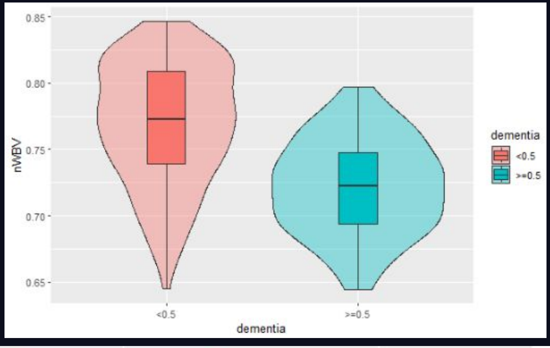
```r
{r}
model3 <- stan_glm(
  dementia_numeric ~ Age + MMSE + nWBV + sex,
  data = oasis_model,
  family = binomial,
  prior_intercept = normal(0, 1.65),
  prior = normal(0, 1, autoscale = TRUE),
  chains = 4,
  iter = 5000*2,
  seed = 84735
)

model3
```

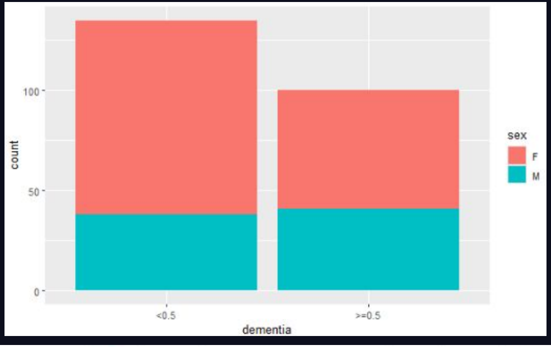
For our project, we chose a **Logistic Regression** model to determine the strongest predictors and interactions of early onset dementia.

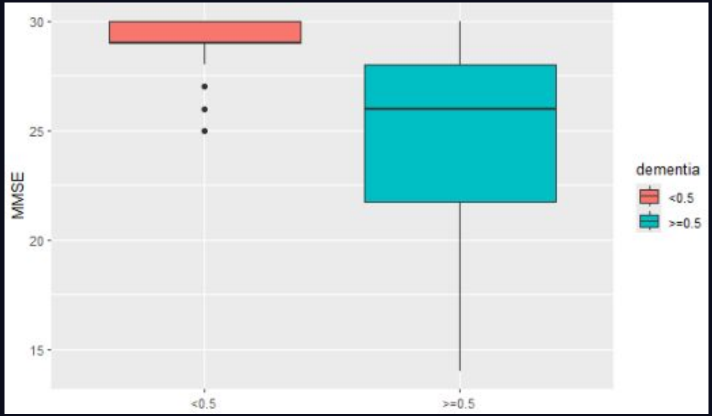
We are using logistic regression because the outcome we are looking for is almost, if this person has/experiences A, B, C; what is the likelihood of early onset dementia?

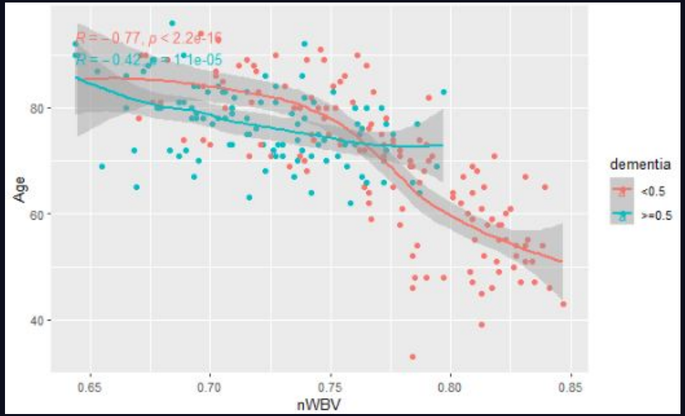
Methods

Before we created our models however, we did some EDA to familiarize ourselves with our data.









Results

```r
{r}
# Model 1: Age + MMSE
model1 <- stan_glm(
  dementia_numeric ~ Age + MMSE,
  data = oasis_model,
  family = binomial,
  prior_intercept = normal(0, 1.65),
  prior = normal(0, 1, autoscale = TRUE),
  chains = 4,
  iter = 5000*2,
  seed = 84735
)

model1
```

	Median	MAD_SD
(Intercept)	6.4	6.2
Age	0.0	0.0
MMSE	0.0	0.2

This model includes Age and MMSE as predictors of dementia. The estimated coefficients are very small and stable, with minimal uncertainty (standard deviations close to 0 for Age and 0.2 for MMSE). The intercept is 6.6 with a standard deviation of 6.2.

```r
{r}
# Model 2: Age + MMSE + nWBV
model2 <- stan_glm(
  dementia_numeric ~ Age + MMSE + nWBV,
  data = oasis_model,
  family = binomial,
  prior_intercept = normal(0, 1.65),
  prior = normal(0, 1, autoscale = TRUE),
  chains = 4,
  iter = 5000*2,
  seed = 84735
)

# Summary
model2
```

	Median	MAD_SD
(Intercept)	7.1	13.2
Age	0.0	0.1
MMSE	0.0	0.2
nWBV	-0.6	14.7

This model adds nWBV as a predictor alongside Age and MMSE. The coefficient for nWBV has a median of -0.6 but a very large standard deviation of 14.7, indicating high uncertainty in its effect. Age and MMSE remain small and stable. The intercept increases slightly to 7.3 (sd 13.1). Overall, adding nWBV introduces more uncertainty into the model coefficients.

```r
{r}
model3 <- stan_glm(
  dementia_numeric ~ Age + MMSE + nWBV + sex,
  data = oasis_model,
  family = binomial,
  prior_intercept = normal(0, 1.65),
  prior = normal(0, 1, autoscale = TRUE),
  chains = 4,
  iter = 5000*2,
  seed = 84735
)

model3
```

	Median	MAD_SD
(Intercept)	7.2	13.1
Age	0.0	0.1
MMSE	0.0	0.2
nWBV	-0.7	14.6
sexM	0.1	1.3

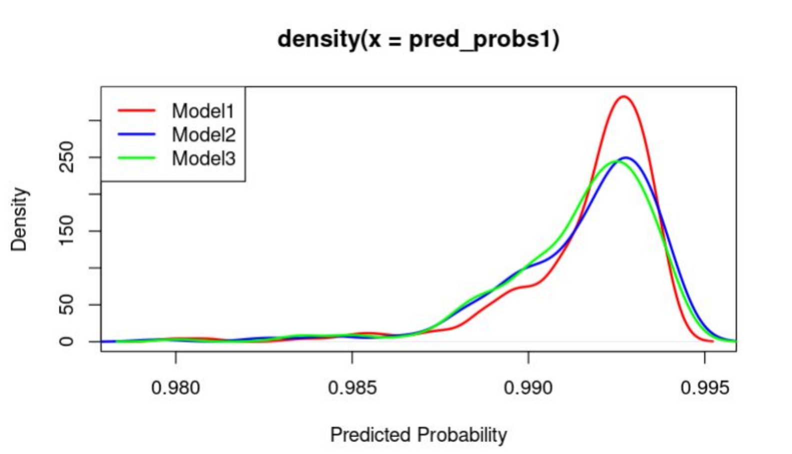
This model further adds `sex` as a predictor in addition to Age, MMSE, and nWBV. The coefficient for sex (`sexM`) is small (0.1) with moderate uncertainty (sd 1.3). The nWBV coefficient remains highly uncertain (median -0.8, sd 14.6). Age and MMSE coefficients are stable as before. The intercept is 7.5 (sd 13.1). Adding sex does not substantially improve the model but slightly increases the number of parameters.

Comparing Models and Variable Impact

Variable	Impact (avg change in predicted probability)
nWBV	0.000801 (~0.08%)
MMSE	0.000967 (~0.097%)

Variable MMSE

Posterior simulations show that changes in MMSE slightly shift dementia probability, while nWBV has almost no effect. MMSE is the best predictor because its effect is consistent and reliable across the dataset, providing meaningful information that nWBV does not.



Model1 (Age + MMSE)

Model 1 proved to be the most practical choice because adding nWBV and sex in Models 2 and 3 only increased uncertainty without improving predictions. By including just Age and MMSE, Model 1 remains simple, easy to interpret, and performs just as well as the more complex models.

Conclusion

The Bayesian analysis showed that Age and Mini Mental State Evaluation (MMSE) are the most important and consistent predictors of dementia in this dataset. Adding other variables like Normalized Whole Brain Volume (nWBV) and sex did not improve predictions, as their effects were uncertain and barely changed the predicted probabilities. Model 1, which includes only Age and MMSE, is the best choice: it is simple, easy to interpret, and performs just as well as the more complex models, highlighting that extra predictors do not provide meaningful benefit.

We expected that lower MMSE and nWBV would both increase dementia risk, with nWBV being especially important in younger people. The results confirmed MMSE as a strong predictor, but nWBV didn't really add anything, and its interaction with age wasn't meaningful. This shows that Age and MMSE explain most of what matters, and the extra factors we thought would help don't actually improve predictions.

Thank you

Repository: https://github.com/etonaki07/DS400_Final_Project/tree/main