PCA

PCA

I needed to remove rows with missing values, which shrunk the dataset from 997 observations to 913. I also needed to center and scale the data.

```
cog = readRDS("./data/cog_data_preproc.RDS") %>% dplyr::select(-subject) %>% na.omit() # omit rows with
cog_pca = cog[, -1] %>% scale # center and scale the data

pca <- prcomp(na.omit(cog_pca))
pca$rotation
pca$sdev
pca$rotation %*% diag(pca$sdev)
corrplot::corrplot(pca$rotation %*% diag(pca$sdev))
```

```
factoextra::fviz_eig(pca, addlabels = TRUE)

factoextra::fviz_contrib(pca, choice = "var", axes = 1)
factoextra::fviz_contrib(pca, choice = "var", axes = 2)
factoextra::fviz_contrib(pca, choice = "var", axes = 3)
factoextra::fviz_contrib(pca, choice = "var", axes = 4)
factoextra::fviz_contrib(pca, choice = "var", axes = 5)
```

Based on the Scree plot, I'd say that the ideal number of PCs is 5. However, in terms of interpretability, I think 3 dimensions is probably sufficient. In the first PC, the white matter volume, cortex volume, cranial volume, and height are the most influential variables. In PC 2, the protective e2 allele and risk e4 allele contribute the most. In PC 3, age and intra-cranial volume. In PC 4, it's weight, height, risk e4 allele, and age. In PC 5, it's the protective e2 allele, risk e4 allele, and weight.

```
factoextra::fviz_pca_biplot(pca,
                            axes = c(1, 2),
                            habillage = cog$cdr,
                            label = c("var"),
                            addEllipses = TRUE)

factoextra::fviz_pca_var(pca, col.var = "steelblue", repel = TRUE)
factoextra::fviz_pca_ind(pca,
                         habillage = cog$cdr,
                         label = "none",
                         addEllipses = TRUE)
```

Above I give plots of associations only for comparing dimensions 1 and 2, but other comparisons for higher dimensions are similar; there is a lot of overlap between the group of patients with and without dimentia, which is our outcome of interest. It looks like it will be difficult to separate them, and we may not be able to fit a highly predictive model as a result.