

Causal Inference for Cognitive Decline Using Mixed Effects Models

Daniel Dema

2025-06-09

Goals

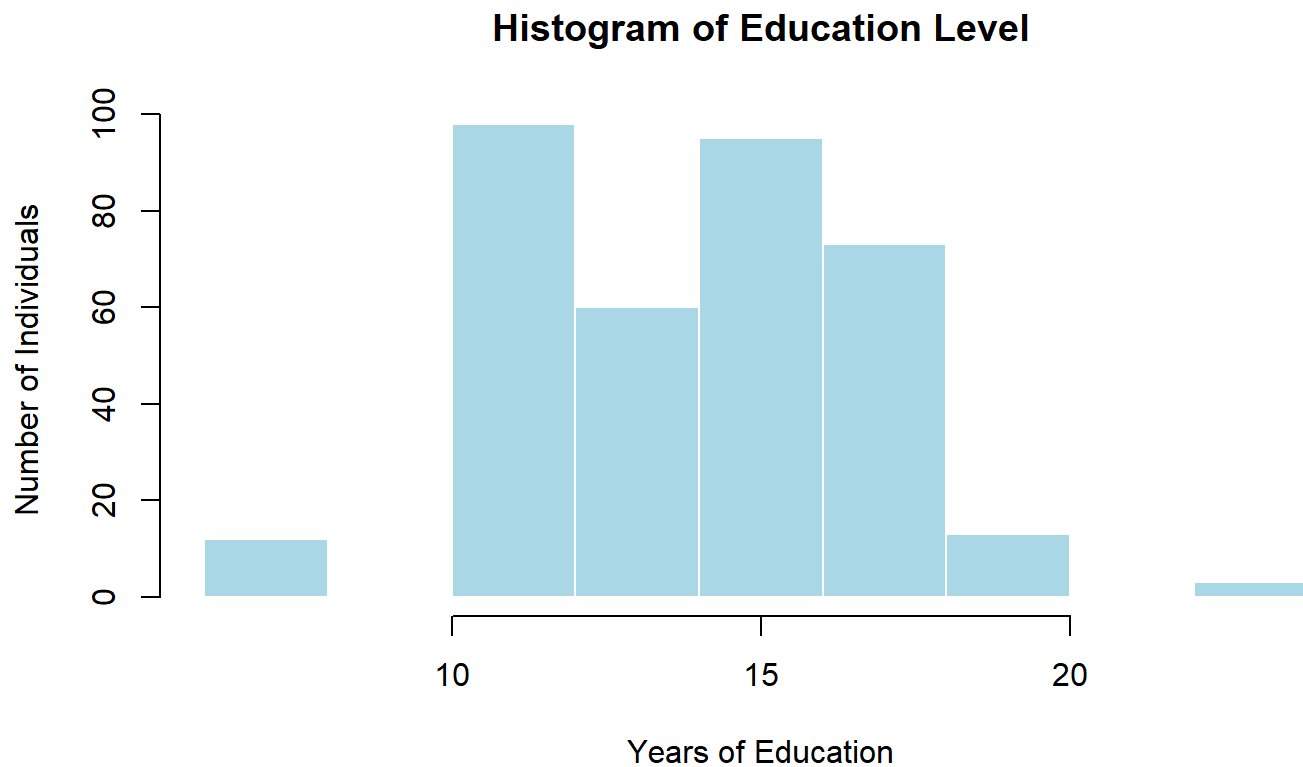
- Predictive inference for cognitive decline in Alzheimer's disease
- Explore the limitations of LMEs on non-normal response data

Data Overview

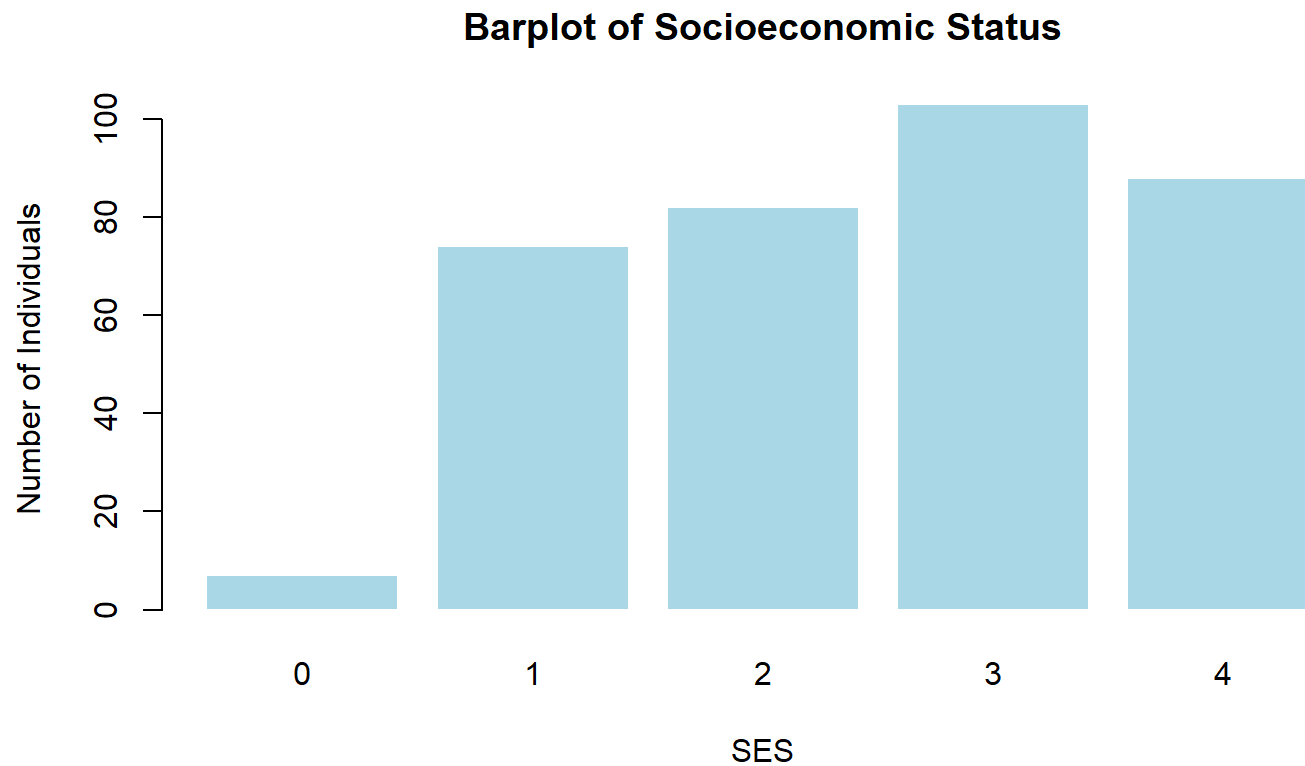
- Subject ID
- Visit number
- Sex
- Age at visit
- Years of education
- Socioeconomic status (SES): score from 0-5
- Normalized whole brain volume (nWBV): ranges from 0-1
- Mini-Mental State Examination (MMSE): test score from 0-30
- Clinical Dementia Ratio (CDR): CDR = 0 (non-demented), CDR = 0.5 (very mild Alzheimer's), CDR = 1 (mild Alzheimer's)

```
df <- read.csv(  
  "D:/Daniel/Documents/MATH6642/final_project/Data/oasis_longitudinal.csv")  
dc <- df[!is.na(df$MMSE) & !is.na(df$SES), ]  
  
dc$SES <- 5 - dc$SES  
  
dc_orig <- dc  
  
dc <- dc[order(dc$Subject.ID, dc$Age), ]
```

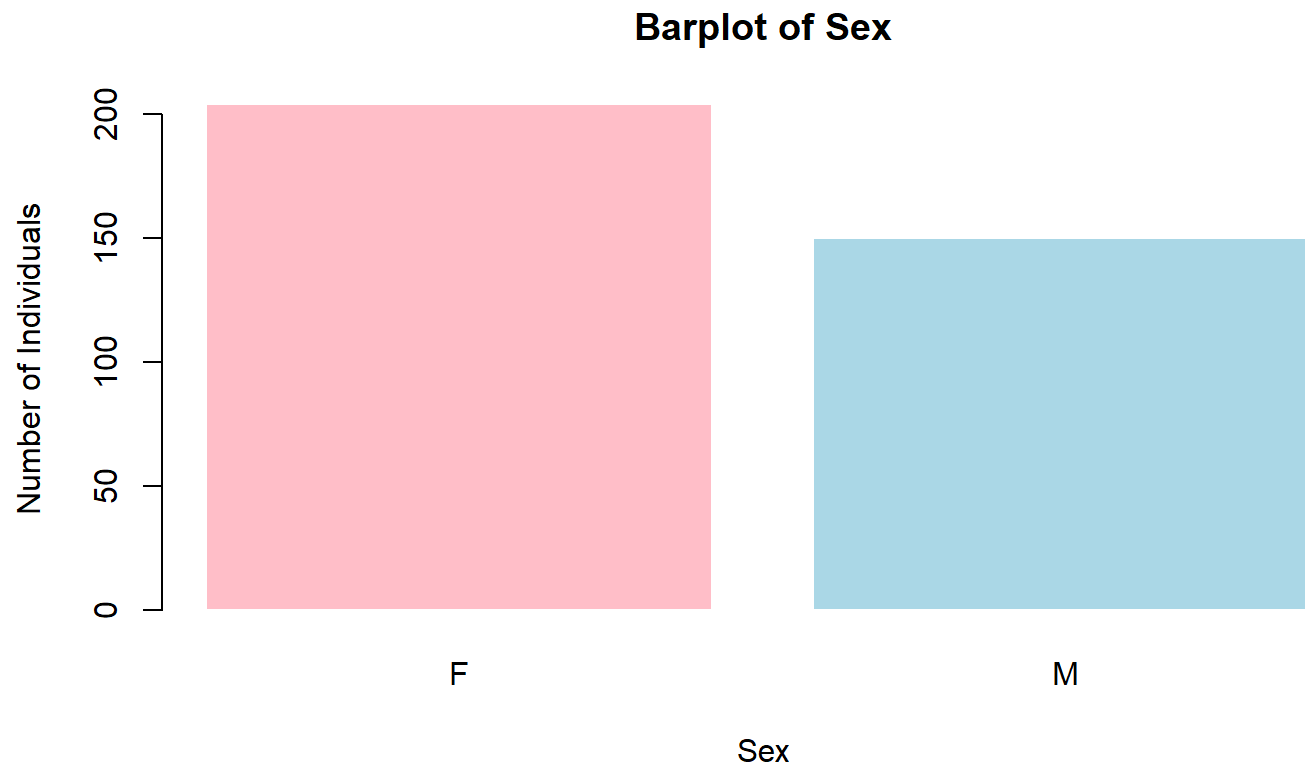
```
hist(dc$EDUC,  
     breaks = 10,  
     main = "Histogram of Education Level",  
     xlab = "Years of Education",  
     ylab = "Number of Individuals",  
     col = "lightblue",  
     border = "white")
```



```
barplot(table(dc$SES),  
  main = "Barplot of Socioeconomic Status",  
  xlab = "SES",  
  ylab = "Number of Individuals",  
  col = "lightblue",  
  border = "white")
```



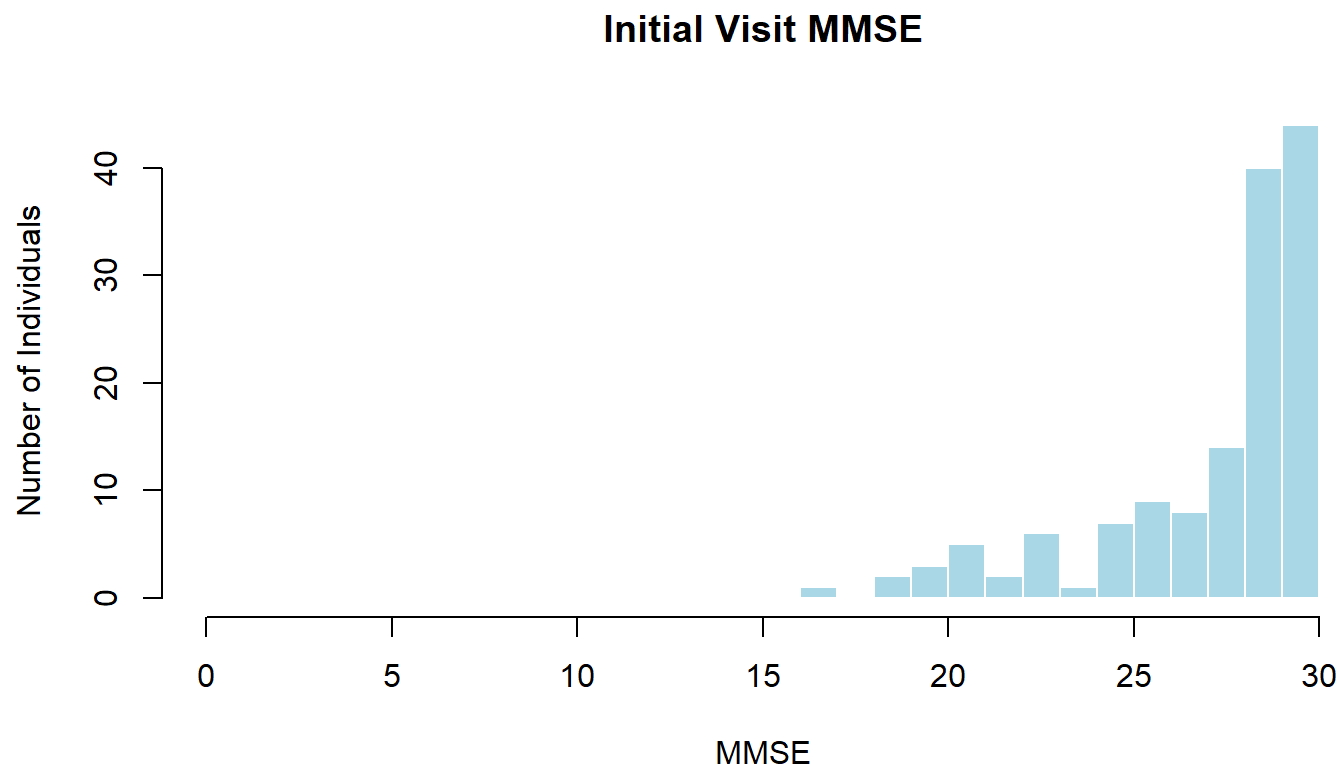
```
barplot(table(dc$M.F),  
        main = "Barplot of Sex",  
        xlab = "Sex",  
        ylab = "Number of Individuals",  
        col = c("pink", "lightblue"),  
        border = "white")
```



```
initial_visits <- dc[ave(dc$Visit, dc$Subject.ID, FUN = min) == dc$Visit, ]  
final_visits <- dc[ave(dc$Visit, dc$Subject.ID, FUN = max) == dc$Visit, ]
```

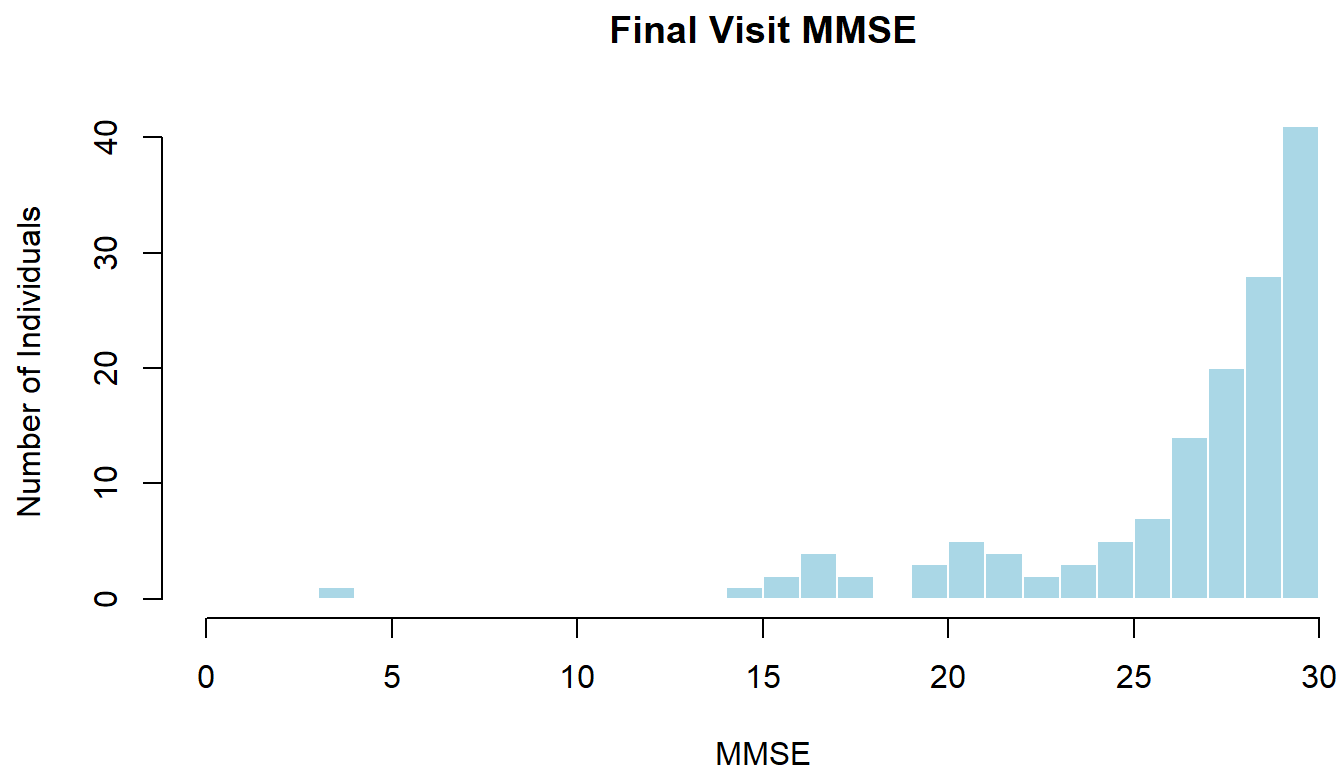
#Histogram for initial visits

```
hist(initial_visits$MMSE, main = "Initial Visit MMSE", breaks = 0:30, xlab = "MMSE",  
      ylab = "Number of Individuals", col = "lightblue", border = "white", xlim = c(0, 30))
```



#Histogram for final visits

```
hist(final_visits$MMSE, main = "Final Visit MMSE", breaks = 0:30, xlab = "MMSE",  
     ylab = "Number of Individuals", col = "lightblue", border = "white", xlim = c(0, 30))
```



Constructing the LME Model

Starting fixed effects:

- Age
- Years of education (EDUC)
- Socioeconomic status (SES)
- Brain volume (nWBV)
- Sex (M.F)

```
fit1 <- lme(MMSE ~ nWBV + Age + M.F + SES + EDUC,  
            data = dc,  
            random = ~ 1 + Age | Subject.ID,  
            correlation = corAR1(form = ~ 1 | Subject.ID))
```

Random effects possibilities:

- $\text{random} = \sim 1 \mid \text{Subject.ID}$
- $\text{random} = \sim 1 + \text{Age} \mid \text{Subject.ID}$
- $\text{random} = \sim 1 + \text{nWBV}$

random = ~ 1 + nWBV has best AIC performance

```
fit2 <- lme(MMSE ~ nWBV + Age + M.F + SES + EDUC,  
  data = dc,  
  random = ~ 1 + nWBV | Subject.ID,  
  correlation = corAR1(form = ~ 1 | Subject.ID),  
  control = lmeControl(opt = "optim", maxIter = 200,  
    msMaxIter = 200))
```

```
VarCorr(fit2)
```

```
## Subject.ID = pdLogChol(1 + nWBV)
##           Variance   StdDev   Corr
## (Intercept) 907.551971 30.125603 (Intr)
## nWBV        1463.567746 38.256604 -1
## Residual     3.542018  1.882025
```

```
summary(fit2)$tTable
```

##	Value	Std.Error	DF	t-value	p-value
## (Intercept)	0.29410337	7.15721747	210	0.04109186	9.672617e-01
## nWBV	29.82370524	7.11650373	210	4.19078053	4.094784e-05
## Age	0.04380276	0.03198838	210	1.36933323	1.723578e-01
## M.FM	-0.66622466	0.44214812	138	-1.50679068	1.341500e-01
## SES	0.09387467	0.26284222	138	0.35715216	7.215235e-01
## EDUC	0.13831876	0.10419206	138	1.32753641	1.865232e-01

For fixed effects:

Possible 3-way interactions:

- nWBV * Age * SES: SES influences access to healthcare, slowing decay in nWBV as the subject ages.
- nWBV * Age * M.F: Possible differences in nWBV decay by sex as the subject ages.
- nWBV * Age * EDUC: Higher education slows the loss of brain volume as the subject ages.


```
fit3 <- lme(MMSE ~ nWBV + Age + M.F + SES + EDUC + nWBV * Age * SES  
  + nWBV * Age + nWBV * SES + Age * SES,  
  data = dc,  
  random = ~ 1 + nWBV | Subject.ID,  
  correlation = corAR1(form = ~ 1 | Subject.ID),  
  control = lmeControl(opt = "optim", maxIter = 200,  
    msMaxIter = 200))
```

```
VarCorr(fit3)
```

```
## Subject.ID = pdLogChol(1 + nWBV)
##           Variance   StdDev   Corr
## (Intercept) 0.03821886 0.1954965 (Intr)
## nWBV        0.04110123 0.2027344 -0.371
## Residual    10.72286737 3.2745790
```

```
summary(fit3)$tTable
```

##	Value	Std.Error	DF	t-value	p-value
## (Intercept)	-109.6451897	93.3145687	206	-1.1750061	0.2413486
## nWBV	159.7594967	127.1639879	206	1.2563266	0.2104204
## Age	1.0239979	1.2219271	206	0.8380188	0.4029914
## M.FM	-0.5032879	0.5390608	138	-0.9336385	0.3521211
## SES	29.7143769	32.3897030	138	0.9174020	0.3605326
## EDUC	0.2038590	0.1294284	138	1.5750715	0.1175297
## nWBV:Age	-1.1008893	1.6808296	206	-0.6549678	0.5132191
## nWBV:SES	-36.1186355	44.4375366	206	-0.8127956	0.4172736
## Age:SES	-0.2651406	0.4224883	206	-0.6275692	0.5309812
## nWBV:Age:SES	0.3072175	0.5846036	206	0.5255143	0.5997909

Let's consider 2-way interactions instead:

- $nWBV * Age$: Brain volume changes with age.
- $nWBV * EDUC$: Education acts as a buffer against brain volume loss.
- $nWBV * SES$: Higher SES allows for better healthcare, improving preservation of brain volume.
- $nWBV * M.F$: Sex affects brain volume.
- $Age * EDUC$: Education acts as a buffer against age effects.
- $Age * M.F$: Aging affects women and men differently.
- $Age * SES$: Higher SES acts as a buffer against age effects by again allowing for better healthcare.

```
fit5 <- lme(MMSE ~ nWBV + Age + SES + nWBV * SES + Age * SES,  
  data = dc,  
  random = ~ 1 + nWBV | Subject.ID,  
  correlation = corAR1(form = ~ 1 | Subject.ID),  
  control = lmeControl(opt = "optim", maxIter = 200,  
    msMaxIter = 200))
```

Lowest AIC so far, all terms statistically significant

```
summary(fit5)$tTable
```

##		Value	Std.Error	DF	t-value	p-value
##	(Intercept)	-54.94655627	16.87724929	208	-3.255658	1.320700e-03
##	nWBV	85.06516352	16.96378182	208	5.014516	1.136780e-06
##	Age	0.24594073	0.07957674	208	3.090611	2.270739e-03
##	SES	21.81791319	6.12511634	140	3.562041	5.034821e-04
##	nWBV:SES	-21.04512536	6.23253893	208	-3.376654	8.755957e-04
##	Age:SES	-0.07694855	0.02853928	208	-2.696233	7.587196e-03

```
VarCorr(fit5)
```

```
## Subject.ID = pdLogChol(1 + nWBV)
##           Variance   StdDev   Corr
## (Intercept) 1025.335434 32.020859 (Intr)
## nWBV        1677.873565 40.961855 -1
## Residual      3.402637  1.844624
```

Diagnostics for G-Matrix

Try within-subject centering?

```
test <- lme(MMSE ~ nWBV + Age + SES + nWBV * SES + Age * SES,  
  data = dc,  
  random = ~ 1 + dvar(nWBV, Subject.ID) | Subject.ID,  
  correlation = corAR1(form = ~ 1 | Subject.ID),  
  control = lmeControl(opt = "optim", maxIter = 200,  
    msMaxIter = 200))
```

```
anova(fit5, test)
```

##	Model	df	AIC	BIC	logLik
## fit5	1	11	1621.422	1663.796	-799.7112
## test	2	11	1643.240	1685.614	-810.6199


```
VarCorr(test)
```

```
## Subject.ID = pdLogChol(1 + dvar(nWBV, Subject.ID))  
##  
##          Variance   StdDev   Corr  
## (Intercept)      7.461877  2.731644 (Intr)  
## dvar(nWBV, Subject.ID) 6328.600137 79.552499 -0.733  
## Residual          2.333263  1.527502
```

```
summary(test)$tTable
```

##		Value	Std.Error	DF	t-value	p-value
##	(Intercept)	-50.99934937	16.98716373	208	-3.002229	3.008278e-03
##	nWBV	79.97829545	17.19428944	208	4.651445	5.857928e-06
##	Age	0.24054110	0.08325873	208	2.889080	4.272828e-03
##	SES	16.51464807	6.21054872	140	2.659129	8.747329e-03
##	nWBV:SES	-15.89944668	6.38933428	208	-2.488436	1.361561e-02
##	Age:SES	-0.05631244	0.02943908	208	-1.912847	5.714120e-02

Trade-off:

Higher AIC, loss of interaction term in exchange for stable G-matrix

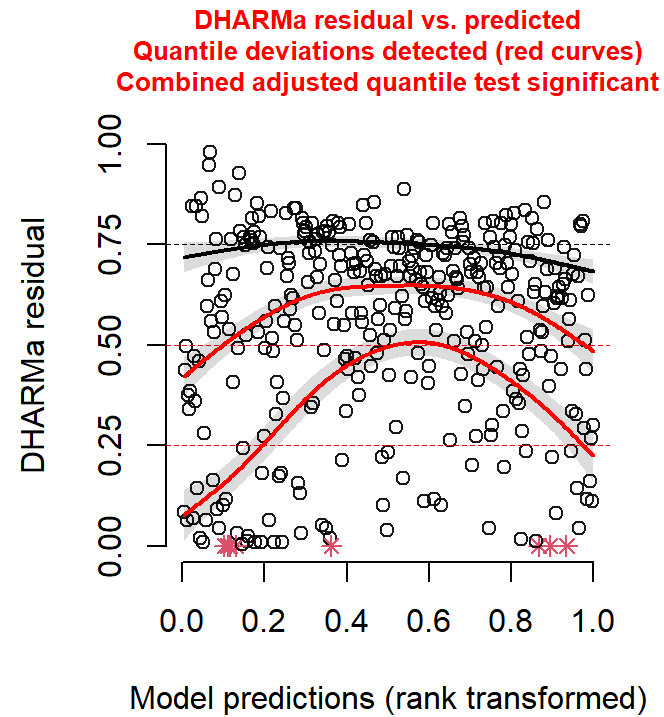
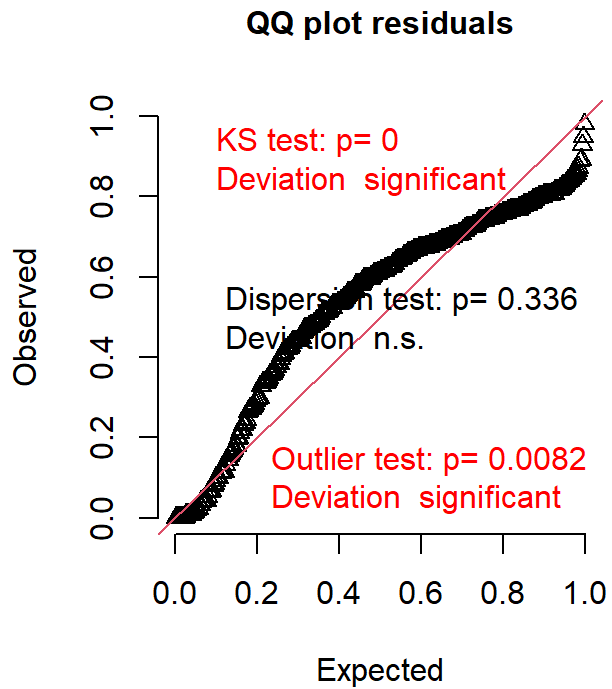
Residuals

Reconstruct fit5 as a glmmTMB so we can use DHARMa to plot residuals:

```
fit5_tmb <- glmmTMB(  
  MMSE ~ nWBV + Age + SES + nWBV:SES + Age:SES + (1 + nWBV | Subject.ID),  
  data = dc,  
  REML = TRUE  
)  
  
res <- simulateResiduals(fittedModel = fit5_tmb)
```

plot(res)

DHARMa residual



Problems:

- Heteroskedasticity
- KS test: residuals are not uniformly distributed
- Outlier test: model handles outliers poorly

Let's try non-linear terms.

```
dc$nwBV_sq <- dc$nwBV^2
```

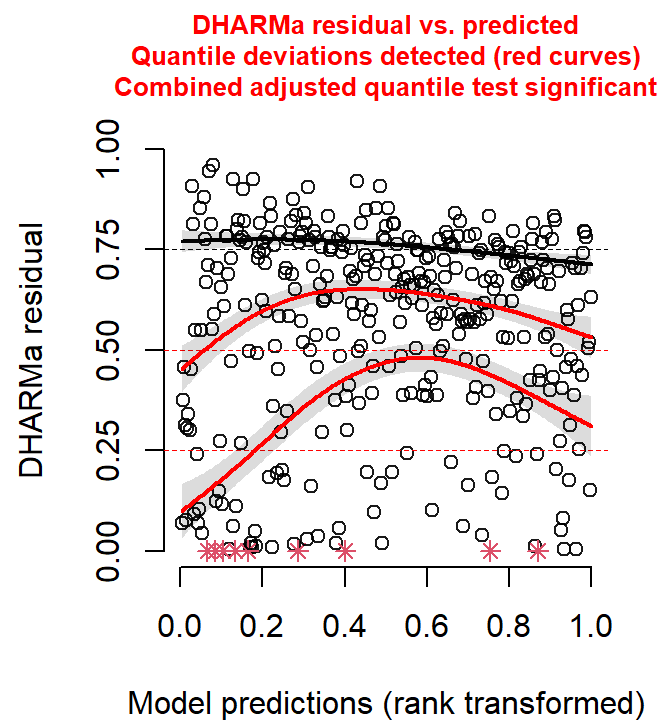
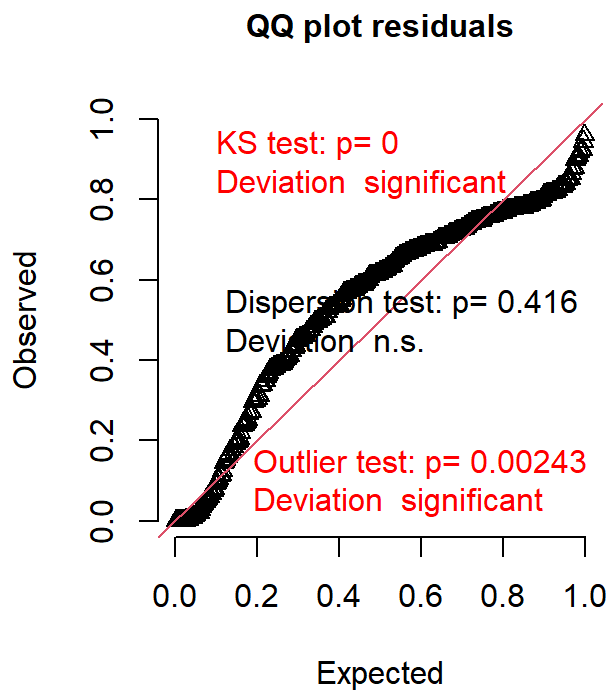
```
fit_nwbv_poly_raw <- glmmTMB(  
  MMSE ~ nwBV + nwBV_sq + Age + SES + nwBV:SES + nwBV_sq:SES + Age:SES  
  + (1 + nwBV | Subject.ID),  
  data = dc,  
  REML = TRUE  
)
```

```
sim_nwbv_poly <- simulateResiduals(fit_nwbv_poly_raw)
```

```
plot(sim_nwbv_poly)
```

```
## Warning in newton(lsp = lsp, X = G$X, y = G$y, Eb = G$Eb, UrS = G$UrS, L = G$L,  
## : Fitting terminated with step failure - check results carefully
```

DHARMA residual




```
AIC(fit_nwbv_poly_raw, fit5_tmb)
```

```
##                df      AIC
## fit_nwbv_poly_raw 12 1601.715
## fit5_tmb          10 1628.256
```

```
dc$nwBV_scaled <- scale(dc$nwBV)[,1]
dc$nwBV_sq_scaled <- dc$nwBV_scaled^2

fit6 <- lmer(MMSE ~ nwBV_scaled + nwBV_sq_scaled + Age + SES +
             nwBV_scaled * SES + Age * SES +
             (1 + nwBV_scaled | Subject.ID),
             data = dc)
```

```
VarCorr(fit6)
```

```
## Groups      Name      Std.Dev. Corr
## Subject.ID (Intercept) 2.3125
##              nWBV_scaled 1.6125  -0.961
## Residual              1.6309
```

```
as.data.frame(summary(fit6)$coefficients[, "Pr(>|t|)"])
```

```
##          summary(fit6)$coefficients[, "Pr(>|t|)"]
## (Intercept)          1.878264e-01
## nWBV_scaled          8.595034e-08
## nWBV_sq_scaled       2.030934e-02
## Age                 1.861178e-03
## SES                 1.169573e-03
## nWBV_scaled:SES      1.926432e-04
## Age:SES              2.587144e-03
```

What if we try splines?

```
dc$ns_nWBV_scaled <- ns(dc$nWBV_scaled, df = 4)
ns_basis <- ns(dc$nWBV_scaled, df = 4)
ns_df <- as.data.frame(ns_basis)
colnames(ns_df) <- paste0("ns_nWBV_", 1:ncol(ns_df))

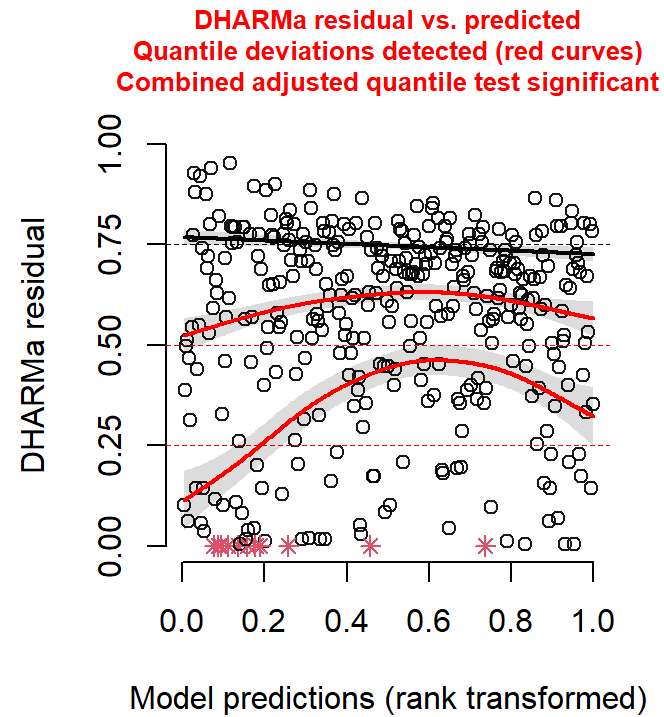
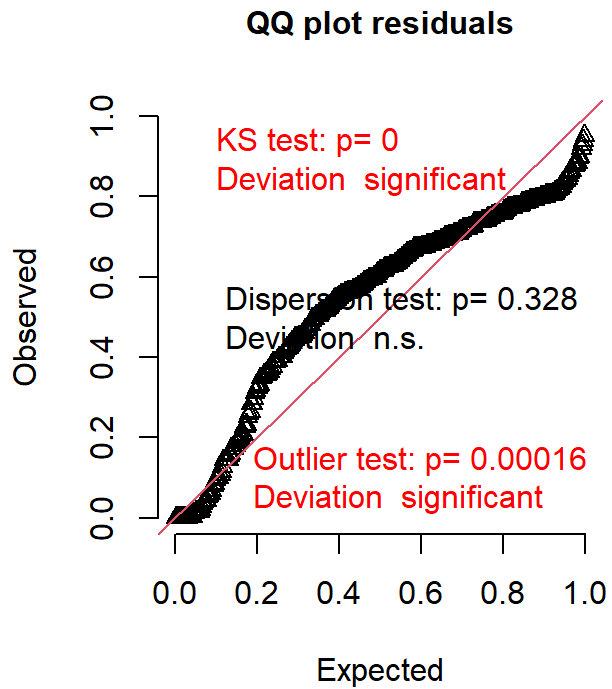
dc <- bind_cols(dc, ns_df)

fit_spline_tmb <- glmmTMB(
  MMSE ~ ns_nWBV_1 + ns_nWBV_2 + ns_nWBV_3 + Age + SES
  + nWBV * SES + Age * SES
  + (1 + nWBV | Subject.ID),
  data = dc,
  REML = TRUE
)

res_spline <- simulateResiduals(fit_spline_tmb)
```

```
plot(res_spline)
```

DHARMa residual



```
dc$nWBV_scaled <- scale(dc$nWBV)[,1]

fit_spline <- lmer(MMSE ~ ns(nWBV_scaled, df=4) + Age + SES +
  nWBV_scaled*SES + Age:SES +
  (1 + nWBV_scaled | Subject.ID),
  data = dc,
  REML = TRUE)

## fixed-effect model matrix is rank deficient so dropping 1 column / coefficient
```

```
VarCorr(fit_spline)
```

```
## Groups      Name      Std.Dev. Corr
## Subject.ID (Intercept) 2.3475
##              nWBV_scaled 1.6483  -0.953
## Residual              1.5943
```



```
as.data.frame(summary(fit_spline)$coefficients[, "Pr(>|t|)"])
```

```
##                                summary(fit_spline)$coefficients[, "Pr(>|t|)"]
## (Intercept)                    3.886609e-01
## ns(nWBV_scaled, df = 4)1        1.144324e-09
## ns(nWBV_scaled, df = 4)2        1.452616e-08
## ns(nWBV_scaled, df = 4)3        1.443120e-09
## ns(nWBV_scaled, df = 4)4        3.127850e-07
## Age                            1.120163e-03
## SES                            9.066069e-04
## SES:nWBV_scaled                1.472280e-04
## Age:SES                        1.991614e-03
```

Solution for Residuals?

Use a glmmTMB with zero-inflation.

```
dc_c <- dc
dc_c$MMSE <- 30 - dc_c$MMSE
dc_c$nWBV_scaled <- scale(dc_c$nWBV)[,1]
dc_c$Age_scaled <- scale(dc_c$Age)[,1]
dc_c$SES_scaled <- scale(dc_c$SES)[,1]

fit5_tmb <- glmmTMB(
  MMSE ~ nWBV_scaled + Age_scaled + SES_scaled +
    nWBV_scaled * SES_scaled + Age_scaled * SES_scaled +
    (1 | Subject.ID),
  data = dc_c,
  ziformula = ~ Age_scaled,
  family = poisson,
  REML = TRUE
)
```

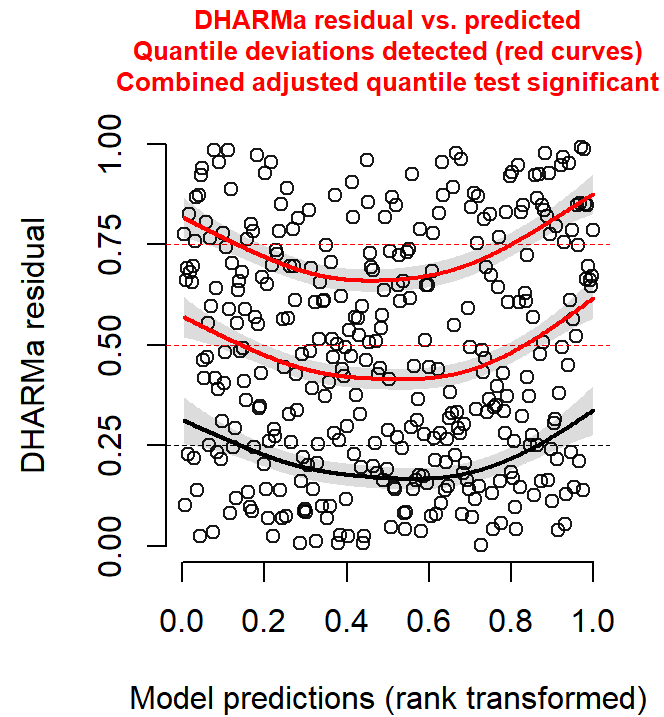
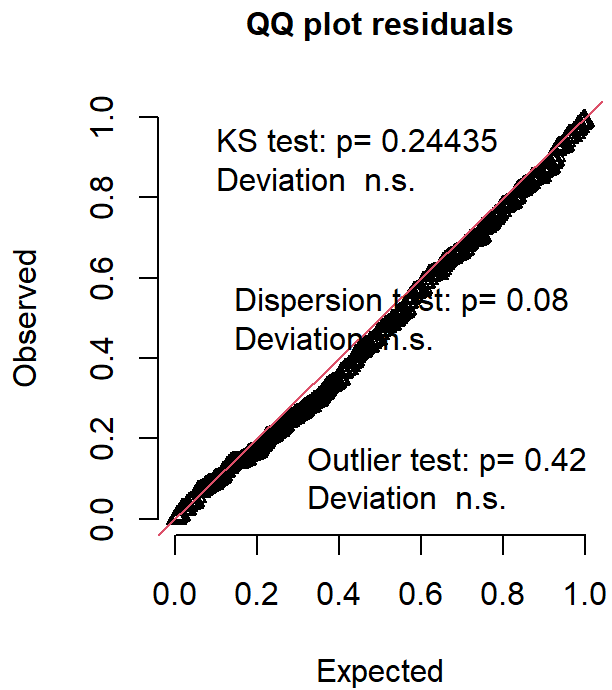
```
AIC(fit5, fit5_tmb)
```

```
## Warning in AIC.default(fit5, fit5_tmb): models are not all fitted to the same  
## number of observations
```

```
##           df      AIC  
## fit5      11 1621.422  
## fit5_tmb   9 1305.129
```

```
res1 <- simulateResiduals(fit5_tmb)
plot(res1)
```

DHARMA residual



Model Interpretation

```
VarCorr(fit_spline)
```

##	Groups	Name	Std.Dev.	Corr
##	Subject.ID	(Intercept)	2.3475	
##		nWBV_scaled	1.6483	-0.953
##	Residual		1.5943	

```
wald(fit_spline)
```

```
## numDF denDF F-value p-value
##      9   Inf 2915.09 <.00001

##              Estimate Std.Error DF  t-value  p-value Lower 0.95
## (Intercept)      -6.066963  7.019796  Inf -0.864265  0.38744 -19.825510
## ns(nWBV_scaled, df = 4)1 12.596798  1.987106  Inf  6.339268 <.00001  8.702142
## ns(nWBV_scaled, df = 4)2 12.740600  2.124430  Inf  5.997184 <.00001  8.576793
## ns(nWBV_scaled, df = 4)3 28.718316  4.562693  Inf  6.294159 <.00001 19.775602
## ns(nWBV_scaled, df = 4)4 16.685178  3.085717  Inf  5.407228 <.00001 10.637283
## Age              0.253806  0.076411  Inf  3.321600  0.00090  0.104044
## SES              7.112171  2.100476  Inf  3.385980  0.00071  2.995314
## SES:nWBV_scaled -0.904538  0.231527  Inf -3.906836  0.00009 -1.358323
## Age:SES         -0.085723  0.027222  Inf -3.149001  0.00164 -0.139077
##              Upper 0.95
## (Intercept)      7.691585
## ns(nWBV_scaled, df = 4)1 16.491454
## ns(nWBV_scaled, df = 4)2 16.904407
## ns(nWBV_scaled, df = 4)3 37.661031
## ns(nWBV_scaled, df = 4)4 22.733073
## Age              0.403568
## SES             11.229029
## SES:nWBV_scaled -0.450754
## Age:SES         -0.032368
```

Model Comparison

```
AIC(fit5, fit6, fit_spline)
```

```
## Warning in AIC.default(fit5, fit6, fit_spline): models are not all fitted to  
## the same number of observations
```

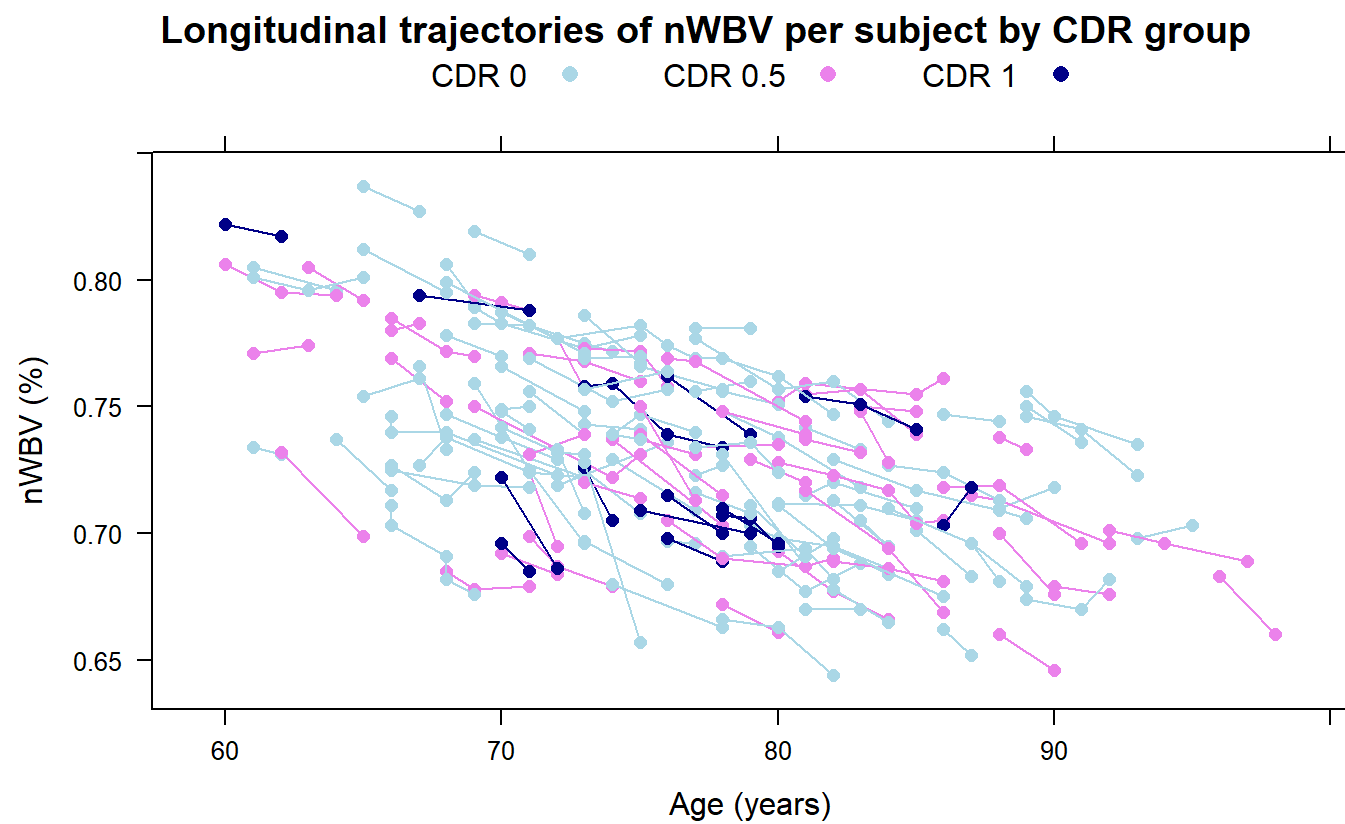
```
##           df      AIC  
## fit5       11 1621.422  
## fit6       11 1639.894  
## fit_spline 13 1624.226
```

```
cdr_colors <- c("0" = "lightblue", "0.5" = "violet", "1" = "darkblue")
```

#nWBV trajectories

```
nWBV_traj <- xyplot(nWBV ~ Age, data = dc,  
  groups = Subject.ID,  
  type = "b",  
  lwd = 1,  
  pch = 16,  
  col = cdr_colors[as.character(dc$CDR)],  
  xlab = "Age (years)",  
  ylab = "nWBV (%)",  
  main = "Longitudinal trajectories of nWBV per subject by CDR group",  
  key = list(text = list(c("CDR 0", "CDR 0.5", "CDR 1")),  
    points = list(pch = 16, col = c("lightblue", "violet", "darkblue")),  
    columns = 3))
```

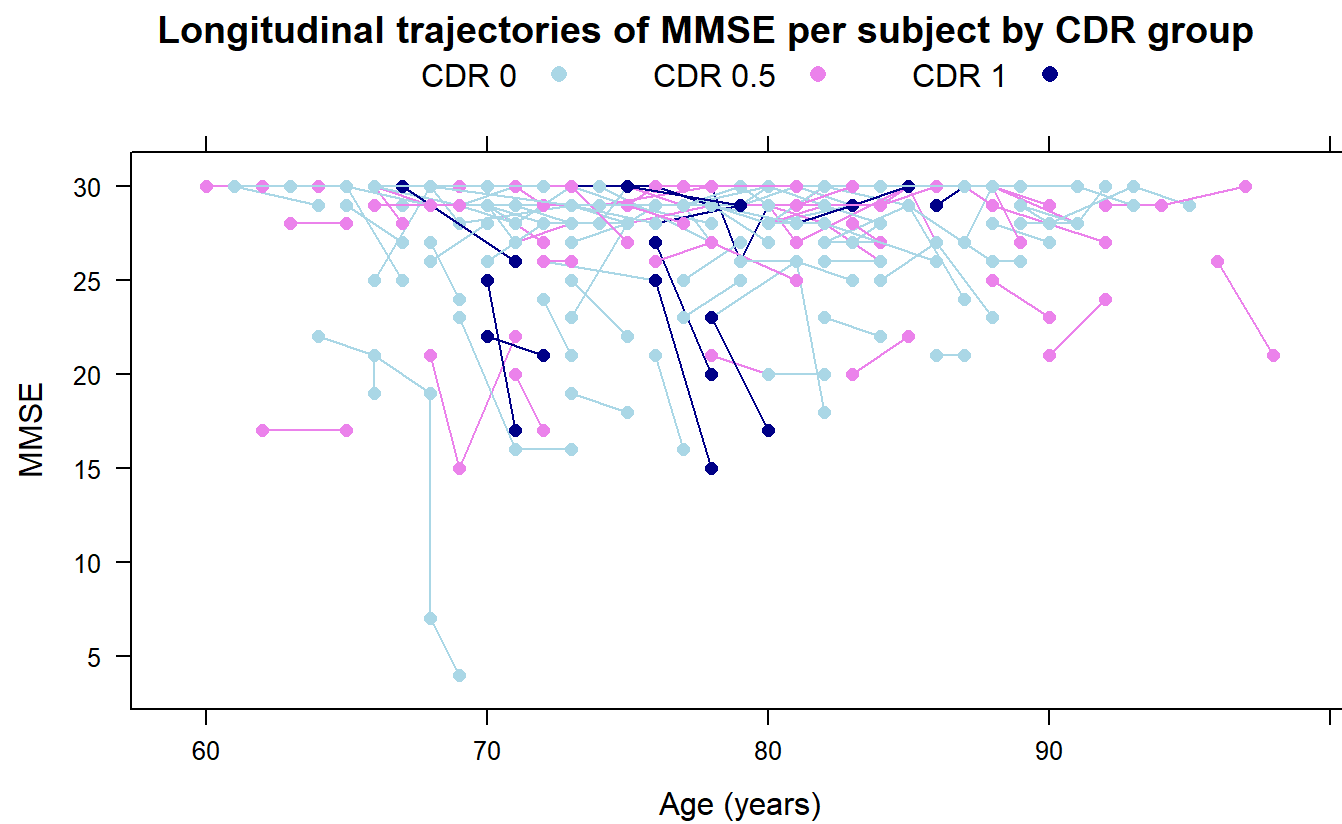

nWBV_traj



#MMSE trajectories

```
MMSE_traj <- xyplot(MMSE ~ Age, data = dc,  
  groups = Subject.ID,  
  type = "b",  
  lwd = 1,  
  pch = 16,  
  col = cdr_colors[as.character(dc$CDR)],  
  xlab = "Age (years)",  
  ylab = "MMSE",  
  main = "Longitudinal trajectories of MMSE per subject by CDR group",  
  key = list(text = list(c("CDR 0", "CDR 0.5", "CDR 1")),  
    points = list(pch = 16, col = c("lightblue", "violet", "darkblue")),  
    columns = 3))
```

MMSE_traj



```
dc_copy <- dc

subject_slopes <- dc_copy %>%
  group_by(Subject.ID, CDR) %>%
  filter(n() > 1) %>% # need at least two points per subject
  summarise(
    slope = {
      fit <- lm(MMSE ~ Age, data = cur_data())
      coef(fit)["Age"]
    },
    .groups = "drop"
  )
```

```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `slope = { ... }`.
## i In group 1: `Subject.ID = "OAS2_0001"` `CDR = 0`.
## Caused by warning:
## ! `cur_data()` was deprecated in dplyr 1.1.0.
## i Please use `pick()` instead.
```

```
avg_slopes_by_CDR_old <- subject_slopes %>%
  group_by(CDR) %>%
  summarise(
    avg_slope = mean(slope, na.rm = TRUE),
    n = n()
```

```
dc$fit_vals5 <- fitted(fit5)
```

```
dc_slope <- dc %>%  
  select(Subject.ID, Age, fit_vals5, CDR) %>%  
  group_by(Subject.ID, CDR) %>%  
  arrange(Age, .by_group = TRUE) %>%  
  summarise(  
    slope = if (n() >= 2) coef(lm(fit_vals5 ~ Age))[2] else NA_real_,  
    .groups = "drop"  
  )
```

```
avg_slopes_by_CDR_5 <- dc_slope %>%  
  group_by(CDR) %>%  
  summarise(  
    avg_slope = mean(slope, na.rm = TRUE),  
    n = n()  
  )
```

```
dc$fit_vals6 <- fitted(fit6)
```

```
dc_slope <- dc %>%  
  select(Subject.ID, Age, fit_vals6, CDR) %>%  
  group_by(Subject.ID, CDR) %>%  
  arrange(Age, .by_group = TRUE) %>%  
  summarise(  
    slope = if (n() >= 2) coef(lm(fit_vals6 ~ Age))[2] else NA_real_,  
    .groups = "drop"  
  )
```

```
avg_slopes_by_CDR_6 <- dc_slope %>%  
  group_by(CDR) %>%  
  summarise(  
    avg_slope = mean(slope, na.rm = TRUE),  
    n = n()  
  )
```

```
dc$fit_spline <- fitted(fit_spline)
```

```
dc_slope <- dc %>%  
  select(Subject.ID, Age, fit_spline, CDR) %>%  
  group_by(Subject.ID, CDR) %>%  
  arrange(Age, .by_group = TRUE) %>%  
  summarise(  
    slope = if (n() >= 2) coef(lm(fit_spline ~ Age))[2] else NA_real_,  
    .groups = "drop"  
  )
```

```
avg_slopes_by_CDR <- dc_slope %>%  
  group_by(CDR) %>%  
  summarise(  
    avg_slope = mean(slope, na.rm = TRUE),  
    n = n()  
  )
```

```
print(avg_slopes_by_CDR_old)
```

```
## # A tibble: 3 × 3
##   CDR avg_slope    n
##   <dbl>   <dbl> <int>
## 1  0     -0.0306   76
## 2  0.5   -0.464    38
## 3  1     -0.826     9
```

```
print(avg_slopes_by_CDR_5)
```

```
## # A tibble: 4 × 3
##   CDR avg_slope    n
##   <dbl>   <dbl> <int>
## 1  0     0.00224   86
## 2  0.5   -0.160    61
## 3  1     -0.946    22
## 4  2    NaN      3
```



```
print(avg_slopes_by_CDR_6)
```

```
## # A tibble: 4 × 3
##   CDR avg_slope    n
##   <dbl>   <dbl> <int>
## 1  0     -0.0135   86
## 2  0.5   -0.227    61
## 3  1     -1.20    22
## 4  2     NaN     3
```

```
print(avg_slopes_by_CDR)
```

```
## # A tibble: 4 × 3
##   CDR avg_slope    n
##   <dbl>   <dbl> <int>
## 1  0     -0.0198   86
## 2  0.5   -0.302    61
## 3  1     -1.29    22
## 4  2     NaN     3
```

- Model with only linear terms is closest for $\text{CDR} = 1$
- Spline model is closest for $\text{CDR} = 0$, $\text{CDR} = 0.5$
- Incorporating non-linearity seems to steepen the mean decline for each CDR group

References

- Marcus, Fotenos, et. al., (2010). *Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults*. MIT Press.
- Marcus, Wang, et al., (2007). *Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults*. MIT Press.
- Boysen, J., (2017). *MRI and Alzheimers*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers>
- Morris, J. C., (1993). *The Clinical Dementia Rating (CDR) : Current version and scoring rules*. Wolters Kluwer.
- Folstein, Folstein & McHugh, (1975). *“Mini-mental state”: A practical method for grading the cognitive state of patients for the clinician*. Elsevier.
- Rasmussen & Langerman, (2019). *Alzheimer’s Disease – Why We Need Early Diagnosis*. Dove Medical Press.

Thanks for your attention!