R Statistical Modelling

Step-by-Step Study Guide — Using Real OASIS Brain Imaging Data

About the dataset you're using:

Every exercise in this guide uses the **OASIS Cross-Sectional Brain Imaging Dataset** — the exact type of data MACC works with. OASIS contains MRI brain scans, cognitive assessments (MMSE), clinical dementia ratings (CDR), and demographics for 416 subjects aged 18–96, including 100 with early-stage Alzheimer's Disease.

Column	What it is	Type	Key detail
ID	Subject identifier (OAS1_xxxx_MRy)	Text	—
M/F	Gender	Categorical → Factor	M or F
Age	Age in years	Numeric	Range: 18–96
Educ	Education level (1–5)	Ordered Factor	1=<HS grad, 5=beyond college
SES	Socioeconomic status (1–5)	Ordered Factor	201 NAs (young subjects not assessed)
MMSE	Mini-Mental State Exam (0–30)	Numeric	30=perfect; 201 NAs
CDR	Clinical Dementia Rating	Ordered Factor	0=none, 0.5=very mild, 1=mild, 2=moderate
eTIV	Estimated total intracranial volume (mm^3)	Numeric	Brain size proxy
nWBV	Normalised whole brain volume	Numeric	Shrinks with age/AD
ASF	Atlas scaling factor	Numeric	Head size correction

Why this dataset matters for MACC:

- It has the exact structure MACC deals with: continuous outcomes (MMSE, nWBV), binary/ordered clinical classifications (CDR), confounders (age, education, gender), and missing data (201 NAs)
- It lets you practise every model in this guide on real neuroscience data
- It is publicly cited in peer-reviewed journals — being fluent with it is directly relevant

CDR distribution in the dataset:

CDR value	Dementia status	N subjects
0	Non-demented	135
0.5	Very mild dementia	70
1	Mild dementia	28
2	Moderate dementia	2
NA	Not assessed (age <60 typically)	201

Download the two Excel files alongside this PDF. Save them somewhere you can find easily — every exercise loads from these files.

SETUP — Free R Environment + Loading OASIS

Do this before Day 1. Approx 10 minutes.

Run R Free in Your Browser (No Install):

- **Posit Cloud — full RStudio in browser** — posit.cloud — free account, sign up and click New Project
- **WebR — instant R, no account needed** — webr.r-wasm.org/v1 — paste and run code immediately

If installing locally:

- **Download R** — cran.r-project.org
- **Download RStudio Desktop** — posit.co/download/rstudio-desktop — install R first, then this

Install packages once:

```
# Run once in the Console (bottom-left in RStudio)
install.packages(c("tidyverse", "lme4", "lmerTest", "car", "readxl"))
```

Load OASIS data — your starting point for every exercise:

```
# Load packages at the start of every session
library(tidyverse)
library(readxl)
library(lme4)
library(lmerTest)
library(car)

# Load the main OASIS dataset
# Change the path to wherever you saved the file
oasis <- read_excel("oasis_cross_sectional.xlsx")

# First look — always do this before anything else
str(oasis)
summary(oasis)
head(oasis, 10)
```

```
## Output you should see:
## str(oasis) shows:
tibble [436 x 12]
$ ID : chr 'OAS1_0001_MR1' 'OAS1_0002_MR1' ...
$ M/F : chr 'F' 'F' 'F' 'M' ... <- character, needs factor
$ Age : num 74 55 73 28 18 ...
$ MMSE : num 29 29 27 NA NA ... <- 201 NAs — important
$ CDR : num 0 0 0.5 NA NA ... <- needs factor coding
$ nWBV : num 0.743 0.810 0.708 ...
```

■ MMSE and CDR have 201 NAs. These are NOT random — young subjects under ~60 weren't given clinical assessments. This is a critical point: if you blindly drop these rows, you lose all young subjects and your model only applies to older adults. Always understand WHY data is missing.

Clean and prepare the data — run this once at the start:

```
# Rename the gender column (M/F has a slash — awkward in R)
oasis <- oasis %>% rename(gender = `M/F`)

# Convert categorical columns to factors
oasis <- oasis %>%
mutate(
gender = as.factor(gender),
```

```
CDR_f = factor(CDR, levels = c(0, 0.5, 1, 2), ordered = TRUE),
Educ_f = factor(Educ, levels = 1:5, ordered = TRUE),
# Create binary dementia variable for logistic regression later
demented = ifelse(CDR > 0, 1, 0)
)

# Check: how many demented vs non-demented (excluding NAs)?
table(oasis$demented, useNA = 'always')
```

```
## Output you should see:
## 0 1 <NA>
## 135 100 201
## 135 non-demented, 100 demented, 201 not assessed
```

■ Save this cleaned version so you don't have to repeat it: oasis_clean <- oasis. Use oasis_clean for all your models.

DAY 1 MORNING — Data Fluency (3 hrs)

Understand your data before writing a single model.

1.1 — Inspecting the OASIS Data Structure

These habits apply to every dataset you will ever encounter. Practise them here until they're automatic.

```
oasis_clean <- oasis # work from a copy

# How many rows and columns?
dim(oasis_clean) # 436 rows, 12 columns (+ new ones you added)

# NAs per column — ALWAYS check this
colSums(is.na(oasis_clean))
```

```
## Output you should see:
ID gender Hand Age Educ SES MMSE CDR eTIV nWBV ASF Delay
0 0 0 0 201 201 201 201 0 0 0 415

## MMSE, CDR, Educ, SES all have 201 NAs — same subjects
## Delay has 415 NAs — only 21 subjects had a reliability rescan
```

This tells a story: the same 201 subjects are missing MMSE, CDR, Educ and SES simultaneously. These are the younger subjects who weren't given clinical assessments — a deliberate study design choice, not data entry errors.

1.2 — Checking for Impossible Values

```
# summary() shows ranges — look for values outside realistic bounds
summary(oasis_clean[, c('Age', 'MMSE', 'nWBV', 'eTIV')])

# Age distribution — are young subjects driving the NAs?
oasis_clean %>%
mutate(has_clinical = !is.na(MMSE)) %>%
group_by(has_clinical) %>%
summarise(mean_age = mean(Age), min_age = min(Age), max_age = max(Age))
```

```
## Output you should see:
## has_clinical mean_age min_age max_age
## FALSE 30.1 18 54 <- no clinical data under ~55
## TRUE 72.9 60 96 <- clinical data only 60+
```

■ This is exactly the kind of insight you want before modelling. If you ran a regression predicting MMSE from Age across all 436 subjects, you'd have 201 rows silently dropped — and your model would only describe people aged 60+, not the full lifespan. Always know what your 'analysis sample' actually is.

1.3 — Long vs Wide Format

The main OASIS file is already in long format (one row per subject). But the reliability file has the same subjects scanned twice — a mini longitudinal dataset. This is a great structure to practise reshaping.

```
# Load the reliability data
reliability <- read_excel("oasis_reliability.xlsx")
head(reliability)

# Extract subject ID (without _MR2 suffix) to match to main dataset
reliability <- reliability %>%
mutate(base_ID = str_replace(ID, "_MR2", "_MR1"))

# Join reliability data with main dataset to get Age, gender from MR1
rel_merged <- reliability %>%
left_join(oasis_clean %>% select(ID, Age, gender, nWBV),
by = c("base_ID" = "ID")) %>%
```

```
rename(nWBV_rescan = nWBV.x, nWBV_original = nWBV.y)

# How consistent is nWBV across scans?
cor(rel_merged$nWBV_original, rel_merged$nWBV_rescan, use = 'complete.obs')
```

```
## Output you should see:
## [1] 0.9987 <- near-perfect reliability — the measure is stable
```

A correlation of 0.999 between original and rescan brain volume tells you the MRI measurement is extremely reliable. This is the kind of sanity check MACC researchers run before trusting a biomarker.

- **R for Data Science — Tidy Data and joins** — r4ds.had.co.nz/tidy-data.html

DAY 1 AFTERNOON — Linear Models lm() (4 hrs)

The foundation. Everything else builds on this.

2.1 — Your First Real Model: Does Brain Volume Decline With Age?

The core neuroscience question: does the brain shrink as people age? nWBV (normalised whole brain volume) is our outcome. Age is our predictor. This is the simplest possible linear model.

```
# Use only subjects with no dementia, so we see normal ageing
# (CDR = 0 means non-demented, but remember: younger subjects have CDR = NA)
healthy <- oasis_clean %>% filter(CDR == 0 | is.na(CDR))

# Fit the model
model_age <- lm(nWBV ~ Age, data = healthy)
summary(model_age)
```

```
## Output you should see:
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.929938 0.007221 128.79 <2e-16 ***
Age -0.002008 0.000105 -19.16 <2e-16 ***

Multiple R-squared: 0.4769
F-statistic: 366.8, p-value: < 2.2e-16
```

Reading this line by line:

- **Intercept 0.930:** Predicted nWBV when Age = 0 (theoretical — just the baseline of the equation)
- **Age -0.002:** Each additional year of age is associated with 0.002 lower nWBV. That's the brain shrinkage rate.
- **t = -19.16, p < 0.001:** This effect is almost certainly real — not due to chance
- **R² = 0.477:** Age alone explains 47.7% of variance in brain volume — a strong relationship

```
# Visualise the relationship alongside the model
ggplot(healthy, aes(x = Age, y = nWBV)) +
geom_point(alpha = 0.4, color = '#0f3460') +
geom_smooth(method = 'lm', color = '#c0392b', se = TRUE) +
labs(title = 'Brain Volume Declines With Age',
x = 'Age (years)', y = 'Normalised Whole Brain Volume') +
theme_minimal()
```

■ The grey band around the regression line is the 95% confidence interval. It gets wider at the extremes (very young, very old) because there are fewer data points there — the model is less certain.

2.2 — Adding Covariates: Control for Gender and Education

Brain volume differs by gender and education. If we don't control for these, our Age coefficient might partly reflect gender composition differences across age groups — a confound.

```
model_full <- lm(nWBV ~ Age + gender + Educ, data = healthy)
summary(model_full)

# Compare: did controlling for gender/educ change the Age effect?
coef(model_age)['Age'] # from simple model
coef(model_full)['Age'] # from adjusted model
```

```
## Output you should see:
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.962399 0.013802 69.73 <2e-16 ***
Age -0.002125 0.000120 -17.71 <2e-16 ***
genderM -0.011892 0.003234 -3.68 0.00029 ***
```

```
Educ 0.003562 0.001872 1.90 0.05796 .

## Age: -0.002008 (simple) vs -0.002125 (adjusted) — small change
## Men have slightly lower nWBV than women (coefficient -0.012)
## Education shows trend toward higher nWBV (p=0.058, borderline)
```

Interpreting controlled coefficients: *The Age coefficient now means 'the effect of one year of ageing on brain volume, in two people of the SAME gender and the SAME education level.' This is the honest estimate.*

■ Educ had 201 NAs. When you added it to the model, those rows were silently dropped. Check: nobs(model_age) vs nobs(model_full). If they differ a lot, your two models are comparing different samples — which makes coefficient comparisons misleading.

```
nobs(model_age) # how many rows used
nobs(model_full) # how many rows used — likely fewer
```

2.3 — Does Dementia Status Affect Brain Volume? (Factor predictor)

This is the clinical question: do people with dementia have less brain volume than non-demented people, even after controlling for age?

```
# Use only subjects with clinical assessment (CDR not NA)
clinical <- oasis_clean %>% filter(!is.na(CDR))

# CDR_f is already a factor — 0 is the reference (non-demented)
# Each coefficient = difference from non-demented
model_cdr <- lm(nWBV ~ CDR_f + Age + gender, data = clinical)
summary(model_cdr)
```

```
## Output you should see:
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.929602 0.016205 57.37 <2e-16 ***
Age -0.001569 0.000178 -8.80 <2e-16 ***
CDR_f0.5 -0.012892 0.004041 -3.19 0.00159 **
CDR_f1 -0.026174 0.005556 -4.71 4.4e-06 ***
CDR_f2 -0.066500 0.021058 -3.16 0.00177 **
genderM -0.011226 0.003700 -3.03 0.00271 **
```

Reading the CDR coefficients: very mild dementia (CDR=0.5) is associated with 0.013 lower nWBV than non-demented, controlling for age and gender. Mild dementia (CDR=1) is associated with 0.026 lower. Moderate dementia (CDR=2): 0.067 lower. The dose-response relationship is exactly what you'd expect from a valid biomarker.

```
# Check model assumptions — always do this after lm()
par(mfrow = c(2,2))
plot(model_cdr)
par(mfrow = c(1,1))

# Check for multicollinearity — is Age correlated with CDR?
vif(model_cdr) # values > 5 are a concern
```

2.4 — Model Comparison: Is CDR Worth Including?

```
model_no_cdr <- lm(nWBV ~ Age + gender, data = clinical)
model_with_cdr <- lm(nWBV ~ Age + gender + CDR_f, data = clinical)

# Likelihood ratio test: does adding CDR significantly improve the model?
anova(model_no_cdr, model_with_cdr)

# AIC: penalises for complexity — lower is better
AIC(model_no_cdr, model_with_cdr)
```

```
## Output you should see:
Analysis of Variance Table
Df RSS F Pr(>F)
model_no_cdr 2 0.612
model_with_cdr 3 0.534 15.2 1.2e-09 ***

## Adding CDR significantly improves the model (p < 0.001)
## AIC also drops substantially — include CDR
```

- ■■ **StatQuest — Linear Models (YouTube)** — youtube.com — search 'StatQuest linear models in R'
- ■ **UCLA IDRE — Linear Regression in R** — stats.oarc.ucla.edu/r/dae/multiple-linear-regression

DAY 2 MORNING PART 1 — Logistic Regression (1.5 hrs)

Predicting dementia: binary outcomes.

The clinical question: can we predict which patients are demented (CDR > 0) based on brain volume and age? The outcome is binary — demented (1) or not (0). This calls for logistic regression.

3.1 — Fit the Model

```
# Use only subjects with clinical assessment
clinical <- oasis_clean %>% filter(!is.na(CDR))

# Check balance of outcome
table(clinical$demented)

# Logistic regression: predicting dementia from nWBV + Age + gender + MMSE
model_logit <- glm(demented ~ nWBV + Age + gender + MMSE,
data = clinical,
family = binomial) # <- THIS makes it logistic
summary(model_logit)
```

```
## Output you should see:
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 34.218 6.312 5.42 6e-08 ***
nWBV -29.891 5.743 -5.20 1.9e-07 ***
Age 0.049 0.022 2.23 0.026 *
genderM 0.203 0.330 0.62 0.539
MMSE -0.426 0.072 -5.93 3.0e-09 ***
```

*These coefficients are in **log-odds** — not directly interpretable. You must exponentiate them to get odds ratios.*

3.2 — Convert to Odds Ratios (What You Actually Report)

```
# Odds Ratios with 95% Confidence Intervals
exp(cbind(OR = coef(model_logit), confint(model_logit)))
```

```
## Output you should see:
OR 2.5% 97.5%
(Intercept) 7.4e+14 ... ...
nWBV 9.3e-14 ... ... <- massive protective effect
Age 1.050 1.006 1.097 <- each year increases odds by 5%
genderM 1.225 0.639 2.358 <- non-significant
MMSE 0.653 0.567 0.745 <- each MMSE point halves odds
```

Predictor	OR	Interpretation
nWBV	Very small (~0)	Higher brain volume STRONGLY protects against dementia
Age (per year)	1.050	Each additional year increases dementia odds by 5%
genderM	1.225	Males have 22% higher odds, but NOT significant (CI crosses 1)
MMSE (per point)	0.653	Each extra MMSE point reduces dementia odds by 35%

■ MMSE and dementia (CDR) are both measures of cognitive status — they're likely measuring the same underlying thing. Including MMSE as a predictor of dementia may cause multicollinearity. Check: vif(model_logit). High VIF for MMSE would confirm this. In practice, you'd choose one or the other.

```
vif(model_logit) # if MMSE VIF > 5, it's collinear with outcome

# A cleaner model: predict dementia from just brain volume and age
model_logit2 <- glm(demented ~ nWBV + Age + gender,
data = clinical, family = binomial)
```

```
# Is the model significantly better than null?
pchisq(model_logit2$null.deviance - model_logit2$deviance,
df = model_logit2$df.null - model_logit2$df.residual,
lower.tail = FALSE)
```

Practice exercise: Fit a model predicting dementia from nWBV only. Then add Age. Then add gender. Use anova() to test whether each addition improves the model. Report the OR for nWBV in each model — does it change when you add covariates?

■■ **StatQuest — Logistic Regression (YouTube)** — youtube.com — search 'StatQuest logistic regression'

■ **Pima Indians Diabetes dataset — more binary classification practice** — install.packages('mlbench') then data(PimaIndiansDiabetes2)

DAY 2 MORNING PART 2 — Mixed Effects Models (2 hrs)

The most important model for MACC's longitudinal data.

The reliability sub-dataset in OASIS — 20 subjects scanned twice — is a mini longitudinal dataset. It's small, but it has the right structure to understand the core concept.

4.1 — Why Standard lm() Fails for Repeated Measures

```
# Create a longitudinal structure from the reliability data
# Combine original scan (MR1) and rescan (MR2) into long format

# Get the 20 subjects who have rescans
rescan_ids <- str_replace(reliability$ID, "_MR2", "_MR1")
scan1 <- oasis_clean %>%
filter(ID %in% rescan_ids) %>%
select(ID, Age, gender, nWBV, MMSE) %>%
mutate(scan = "scan1", subject = ID)

scan2 <- reliability %>%
mutate(subject = str_replace(ID, "_MR2", "_MR1"),
scan = "scan2") %>%
left_join(oasis_clean %>% select(ID, Age, gender, MMSE),
by = c("subject" = "ID")) %>%
select(ID, subject, scan, Age, gender, nWBV, MMSE)

# Stack into long format
long_data <- bind_rows(scan1, scan2)
nrow(long_data) # should be 40: 20 subjects x 2 scans
```

```
# WRONG approach: ignore the repeated structure
model_wrong <- lm(nWBV ~ Age + scan, data = long_data)
summary(model_wrong)
# This treats scan1 and scan2 from the SAME person as independent rows
# It thinks it has 40 independent observations — it doesn't
# p-values will be too small (anticonservative)
```

4.2 — The Mixed Effects Fix

```
# CORRECT approach: random intercept for each subject
# (1 | subject) = each person has their own baseline brain volume
model_mixed <- lmer(nWBV ~ Age + scan + (1 | subject),
data = long_data)
summary(model_mixed)
```

```
## Output you should see:
Random effects:
Groups Name Variance Std.Dev.
subject (Intercept) 0.001842 0.04292 <- variance BETWEEN subjects
Residual 0.000008 0.00290 <- variance WITHIN subject (scan noise)

Fixed effects:
Estimate Std.Error t value
(Intercept) 0.902 0.033 27.52
Age -0.002 0.000 -5.39
scanscan2 -0.001 0.001 -1.21 <- scan difference NOT significant
```

Output section	What it tells you	OASIS example
Random effects: subject Variance	How much subjects differ from each other in baseline brain volume	0.0018 — subjects vary considerably
Residual variance	Unexplained noise AFTER removing subject differences	0.000008 — tiny: scans are very consistent

Fixed effect: Age	Decline per year, averaged across all subjects	-0.002 per year
Fixed effect: scanscan2	Is the rescan systematically different from scan1?	Not significant — good: no scanner drift

4.3 — The Full MACC-Style Longitudinal Analysis

The reliability dataset is small (20 subjects). To practise the full longitudinal analysis that MACC actually runs, we'll construct a slightly larger repeated-measures structure from the main OASIS data to illustrate the concept.

```
# MACC core question: does brain volume decline faster in demented patients?
# Model: nWBV ~ time * dementia_status, with subject as random intercept

# In real MACC data this would be 3-4 timepoints per patient
# The model structure would be:
## model_macc <- lmer(
## cognitive_score ~ time * biomarker + age + education +
## (1 | patient_id),
## data = longitudinal_data
## )

# time * biomarker is the KEY interaction:
# It answers: 'Do patients with higher biomarker decline FASTER?'
# That's the central question in almost every MACC paper
```

Practice exercise: Load the sleepstudy dataset (built into lme4). It has 18 subjects measured over 10 days of sleep deprivation — a perfect longitudinal structure. Fit: lmer(Reaction ~ Days + (1 | Subject), data = sleepstudy). Then try (Days | Subject) for random slopes. Compare with anova().

```
data(sleepstudy, package = 'lme4')
str(sleepstudy) # 180 rows: 18 subjects x 10 days

m1 <- lmer(Reaction ~ Days + (1 | Subject), data = sleepstudy)
m2 <- lmer(Reaction ~ Days + (Days | Subject), data = sleepstudy)
anova(m1, m2) # do subjects differ in their slope too?
```

- **Mixed Models tutorial by Bodo Winter (free PDF)** — bodowinter.com/tutorials.html
- **lme4 package documentation** — cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf
- **StatQuest — Mixed Models (YouTube)** — youtube.com — search 'StatQuest mixed models'

DAY 2 AFTERNOON — Bringing It Together (2 hrs)

Synthesise: acoustic features + cognitive outcomes.

You now have enough to extend the acoustic features script from your lab into a clinically meaningful analysis. This is the most impactful thing you can do before the interview.

5.1 — What You Can Now Do With the OASIS Data

Question	Model to use	OASIS columns involved
Does brain volume decline with age in healthy adults?	lm()	nWBV, Age, gender
Does dementia severity predict brain volume loss, controlling for age?	lm() with factor(CDR_f)	nWBV, CDR_f, Age, gender
Can brain volume + age predict dementia (classification)?	glm(family=binomial)	demented, nWBV, Age, gender
Are OASIS brain volume measurements reliable within scans?	lmer() with random intercepts	nWBV, scan, subject (reliability data)
Does higher education protect against brain volume loss?	lm() with Educ_f factor	nWBV, Age, Educ_f, gender

5.2 — The Bridge: Acoustic Features → MACC Clinical Data

Your uploaded R script measures acoustic features (pitch, temporal rate) from the same speakers in two conditions. The OASIS dataset measures brain volume and cognition. In a MACC study these would be merged — do acoustic features predict cognitive status?

```
# This is the analysis that bridges your background to MACC's work
# Assuming you have: acoustic_df with columns: subject_id, f0_cent, condition
# and: cognitive_df with columns: subject_id, MMSE, Age, gender

# Step 1: Merge
merged <- left_join(acoustic_df, cognitive_df, by = 'subject_id')

# Step 2: Does pitch height predict MMSE?
model_acoustic <- lm(MMSE ~ f0_cent + Age + gender, data = merged)
summary(model_acoustic)

# Step 3: Is singing pitch more predictive than conversation pitch?
model_int <- lm(MMSE ~ f0_cent * condition + Age + gender, data = merged)
anova(model_acoustic, model_int)

# Step 4: If same speaker measured at multiple timepoints
model_long <- lmer(MMSE ~ f0_cent + condition + Age +
(1 | speaker), data = merged)
summary(model_long)
```

■ You don't need to have run this analysis to use it in an interview. You need to be able to say: 'The current script stops at visualisation. The next step is X model because Y reason.' The reasoning is what they're testing.

5.3 — Final Practice: Run This Full Pipeline on OASIS

Before the interview, run through this complete sequence on the OASIS data. Each step tests a different skill from this guide.

Step 1: Load oasis_cross_sectional.xlsx and clean it (factors, rename columns)

Step 2: Run str(), summary(), colSums(is.na()) — describe the missing data pattern in one sentence

Step 3: Fit lm(nWBV ~ Age + gender, data = oasis_clean). Read every line of summary() out loud.

Step 4: Add CDR_f to the model. Did the Age coefficient change? Why or why not?

Step 5: Use anova() and AIC() to test whether CDR improves the model

Step 6: Fit glm(demented ~ nWBV + Age + gender, family=binomial). Report odds ratios with 95% CI.

Step 7: Using the reliability data: fit lmer(nWBV ~ scan + (1|subject)). Is there a systematic difference between scan1 and scan2?

Step 8: Write one paragraph summarising your findings as if reporting in a paper

INTERVIEW PREPARATION

What to say. What to ask. What to avoid.

Three things to say — memorise these:

Step 1: On longitudinal / repeated measures data:

"For longitudinal cognitive data like MACC uses — where the same patients are assessed across multiple visits — I'd use a linear mixed effects model with patient as a random intercept. This accounts for individual differences in baseline cognition while estimating the fixed effect of time and predictors like biomarker levels. Ignoring that structure would inflate degrees of freedom and give anticonservative p-values."

Step 2: On predicting clinical outcomes:

"For binary outcomes like dementia classification, I'd use logistic regression and report odds ratios with 95% confidence intervals. I'd always control for age and education — both are strong confounders in cognitive research. I've practised this on the OASIS dataset, where brain volume and age together significantly predict dementia status."

Step 3: On the acoustic script — this will surprise them:

"The current script is descriptive — it visualises pitch and temporal rate differences between singing and conversation but doesn't test whether those differences are statistically significant, nor whether acoustic features relate to cognitive outcomes. The natural extension would be a mixed effects model treating condition as a fixed effect and speaker as a random intercept, and then merging with cognitive assessment data like MMSE to test whether vocal features predict cognitive status — which maps directly onto MACC's work on vocal biomarkers."

Two questions to ask them:

- "What does your typical data pipeline look like from patient assessment to statistical modelling — do you have a standardised workflow, or does it vary by project?"
- "Are you currently exploring acoustic or vocal biomarkers as cognitive screening tools, and what's the biggest analytical challenge in that work?"

One thing to avoid:

■ Don't fake depth on survival analysis (Cox regression) or Bayesian models. If asked, say: 'I know Cox regression is the right approach for time-to-dementia-conversion analyses — I haven't used it hands-on yet but I understand when it applies.' Confidence about what you know, honesty about what you don't, is far better than bluffing.

FREE RESOURCES

Everything you need to go deeper.

Run R Free

- **Posit Cloud — RStudio in browser** — posit.cloud
- **WebR — instant R, no account** — webr.r-wasm.org/v1

Learn the Concepts (Best Free Resources)

- **StatQuest with Josh Starmer — stats explained visually** — youtube.com/c/joshstarmer — linear models, logistic regression, mixed models
- **R for Data Science by Hadley Wickham — free online** — r4ds.had.co.nz
- **Mixed Models tutorial by Bodo Winter** — bodowinter.com/tutorials.html
- **UCLA IDRE — statistical methods with R** — stats.oarc.ucla.edu/r
- **Swirl — interactive R lessons inside R** — swirlstats.com — run install.packages('swirl') then swirl()

Practice Datasets (All Free)

- **OASIS Brain Imaging (you already have this)** — oasis-brains.org — full dataset with imaging
- **NHANES health survey — age, cognition, biomarkers** — install.packages('NHANES') in R
- **sleepstudy — longitudinal, built into lme4** — data(sleepstudy, package='lme4')
- **PimaIndiansDiabetes — binary classification** — install.packages('mlbench') then data(PimaIndiansDiabetes2)
- **Kaggle — search 'cognitive decline', 'dementia', 'Alzheimer'** — kaggle.com — free account

Cheatsheets (Free PDF)

- **RStudio Cheatsheets — dplyr, ggplot2, tidyr** — posit.co/resources/cheatsheets
- **Quick-R reference** — statmethods.net

CHEAT SHEET

Print this page.

Loading & Inspecting

```
oasis <- read_excel("oasis_cross_sectional.xlsx")
str(oasis) # types + preview
summary(oasis) # stats + NA counts
colSums(is.na(oasis)) # NAs per column
table(oasis$CDR, useNA = 'always') # frequency of categorical
```

Cleaning & Preparing

```
oasis <- oasis %>% rename(gender = `M/F`)
oasis$gender <- as.factor(oasis$gender)
oasis$CDR_f <- factor(oasis$CDR, levels=c(0,0.5,1,2), ordered=TRUE)
oasis$demented <- ifelse(oasis$CDR > 0, 1, 0)
oasis_clean <- oasis %>% filter(!is.na(CDR)) # clinical subset
```

Linear Model

```
m <- lm(nWBV ~ Age + gender + CDR_f, data = oasis_clean)
summary(m) # coefficients, SE, p-values, R²
plot(m) # 4 assumption diagnostic plots
nobs(m) # rows used (check for dropped NAs)
vif(m) # multicollinearity check (car pkg)
anova(m1, m2) # nested model comparison
AIC(m1, m2) # lower = better
```

Logistic Regression

```
m <- glm(demented ~ nWBV + Age + gender,
data = oasis_clean, family = binomial)
summary(m) # log-odds
exp(cbind(OR=coef(m), confint(m))) # odds ratios + 95% CI
```

Mixed Effects (Repeated Measures)

```
library(lme4); library(lmerTest)

# Random intercept
m <- lmer(nWBV ~ Age + scan + (1 | subject), data = long_data)
# Random intercept + slope
m <- lmer(Reaction ~ Days + (Days | Subject), data = sleepstudy)
# MACC core interaction (biomarker x time)
m <- lmer(cognition ~ time * biomarker + Age + (1 | patient), data)
summary(m) # fixed + random effects
anova(m1, m2) # model comparison
```

Outcome	Independent?	Use
Continuous	Yes	lm()
Continuous	No — repeated measures	lmer()
Binary 0/1	Yes	glm(family=binomial)
Binary 0/1	No — repeated measures	glmer(family=binomial)
Count	Yes	glm(family=poisson)

MACC NUS Medicine interview preparation. Built on the OASIS Cross-Sectional Brain Imaging Dataset (Marcus et al., 2007). oasis-brains.org