Selecting a Working Correlation Structure in GEE:

Methods, Simulation, and Example with Simulated FEV1 Data

March 13, 2025

1 Introduction

Correlated responses are common in biomedical studies. A typical example is the *longitudinal study*, where each subject is followed over multiple time points and repeated observations of the response variable (e.g., forced expiratory volume in one second, FEV1) and relevant covariates (e.g., smoking status) are recorded. Since repeated observations come from the same individual, standard generalized linear models (GLMs) that assume independence may not be appropriate.

The **generalized estimating equation (GEE)** approach (Liang and Zeger) provides a way to handle such correlated data by specifying a working correlation structure. Examples include *independence*, *compound symmetry (CS)*, or *first-order autoregressive (AR-1)*. While GEE estimates remain consistent under mild assumptions, the choice of correlation structure can affect efficiency (variance) and, in some cases, bias — especially with time-varying covariates (**??**).

To systematically choose among candidate correlation structures, one can compare models using **predictive mean squared error (PMSE)**, which measures a model's average prediction error on new data. Since additional independent data are typically unavailable, **resampling methods** such as the bootstrap or cross-validation are used to approximate PMSE in practice (**??**).

2 Simulation: Generating FEV1 Data

Below is an example of how to **simulate** a longitudinal dataset for `N_all` = 1000 subjects, each measured over `T_time` = 5 years. We incorporate baseline covariates, a random intercept, and a time-varying smoking status.

2.1 R Code for Data Generation

```
N_all <- 1000        # Total sample size
T_time <- 5          # Number of time points (5 years)
sigma_b <- 0.8       # SD of random intercepts
sigma_e <- 1.0       # SD of measurement error

# Regression coefficients (beta)
beta0 <- 2.4
beta_smoking <- -0.15
beta_treat <- 0.10
beta_age <- -0.02
beta_gender <- 0.30
beta_BMI <- -0.03
beta_weight <- 0.01
beta_packyears <- -0.02
beta_cigsday <- -0.005
beta_sbp <- -0.004

# 1. Generate baseline covariates
id <- 1:N_all
age <- rnorm(N_all, mean = 50, sd = 10)
gender <- rbinom(N_all, 1, 0.5)
BMI <- rnorm(N_all, mean = 25, sd = 3)
weight <- rnorm(N_all, mean = 70, sd = 10)
pack_years <- rpois(N_all, lambda = 20)
cigs_per_day <- rpois(N_all, lambda = 15)
sbp <- rnorm(N_all, mean = 120, sd = 15)
treatment <- rbinom(N_all, 1, 0.5)

b_i <- rnorm(N_all, mean = 0, sd = sigma_b)  # Random intercept
```

```
df_subjects <- data.frame(
  id, age, gender, BMI, weight,
  pack_years, cigs_per_day, sbp,
  treatment, b_i
)

# 2. Generate repeated measurements
dat_all <- data.frame()
for(i in seq_len(N_all)){
  for(t in 1:T_time){
    # Time-varying smoking status: 40% chance of smoking each year
    smoking_it <- rbinom(1, 1, 0.4)

    # Linear predictor
    mu <- beta0 +
          beta_smoking     * smoking_it +
          beta_treat       * df_subjects$treatment[i] +
          beta_age         * df_subjects$age[i] +
          beta_gender      * df_subjects$gender[i] +
          beta_BMI         * df_subjects$BMI[i] +
          beta_weight      * df_subjects$weight[i] +
          beta_packyears   * df_subjects$pack_years[i] +
          beta_cigsday     * df_subjects$cigs_per_day[i] +
          beta_sbp         * df_subjects$sbp[i] +
          df_subjects$b_i[i]   # Random intercept

    # Generate FEV1 with measurement error
    FEV1_it <- rnorm(1, mean = mu, sd = sigma_e)

    dat_all <- rbind(dat_all, data.frame(
      id = i,
      time = t,
      smoking = smoking_it,
      treatment = df_subjects$treatment[i],
      age = df_subjects$age[i],
```

```
      gender = df_subjects$gender[i],
      BMI = df_subjects$BMI[i],
      weight = df_subjects$weight[i],
      pack_years = df_subjects$pack_years[i],
      cigs_per_day = df_subjects$cigs_per_day[i],
      sbp = df_subjects$sbp[i],
      FEV1 = FEV1_it
    ))
  }
}
```

In this simulation, each participant i has:

- **Baseline covariates**: age, gender, BMI, etc.
- A **random intercept** b_i.
- A **time-varying smoking status** (40% chance of smoking each year).
- An outcome $FEV1_{it}$ influenced by both baseline and time-varying factors, plus measurement error $\epsilon \sim N(0, \sigma_e^2)$.

3 Cross-sectional Analysis and Preliminary Results

To illustrate the difference between a purely cross-sectional approach and a longitudinal model, one might fit a separate linear model for each time point (`Year` $= 1, \ldots, 5$):

```
fit <- lm(FEV1 ~ smoking + treatment + age + gender + BMI + weight +
                 pack_years + cigs_per_day + sbp,
          data = df_year)
```

Below is an example of the smoking coefficient estimates (`Smoking_est`) and their standard errors (`Smoking_se`) across the five years:

In our simulation, we generated three data sets of different sizes: $n = 100$, $n = 500$, and $n = 1000$, each containing 5 years of measurements. For each data set, we fit a separate linear regression model at each of the 5 time points (Year 1 to Year 5). Table 2 presents the estimated coefficient (`Smoking_est`) and its standard error (`Smoking_se`) for the current-year smoking status across the five years in each of the three sample sizes.

Year	Smoking_est	Smoking_se
1	-0.5186916	0.3262590
2	-0.1440486	0.2606193
3	-0.0751477	0.2542296
4	-0.1468116	0.2572298
5	-0.1462442	0.2828489

Table 1: Cross-sectional regression estimates for smoking at each year.

As the sample size increases from $n = 100$ to $n = 1000$, we typically expect the standard errors to decrease due to more data being available. However, the actual point estimates (`Smoking_est`) may vary depending on random fluctuations in the simulated data. In some cases, the estimated effect of smoking could be close to zero (or even positive in certain runs), reflecting the stochastic nature of simulation and the chosen parameter values.

In some runs, the p-values for smoking or other covariates may not show strong significance, partly because the data were simulated with certain parameter settings and random effects. For a true longitudinal analysis, we would typically employ GEE or mixed-effects models to account for within-subject correlation.

4 GEE and Working Correlation Structures

4.1 Independence, CS, and AR-1

Once we have the simulated data (or real data), we can fit a **GEE model** using different correlation structures:

- **Independence**: assumes no within-subject correlation.
- **Compound Symmetry (CS)**: assumes a common correlation ρ between any two time points for the same subject.
- **AR-1 (First-order Autoregressive)**: assumes correlation decays with time-lag $|t_1 - t_2|$.

Year	$n = 100$		$n = 500$	
	Smoking_est	Smoking_se	Smoking_est	Smoking_se
1	-0.5186916	0.3262590	0.03555778	0.1226416
2	-0.1440486	0.2606193	-0.4144076	0.2116636
3	-0.0751477	0.2542296	0.01104522	0.1683850
4	-0.1468116	0.2572298	-0.1143029	0.2151824
5	-0.1462442	0.2828489	-0.06680580	0.0843020

Year	$n = 1000$	
	Smoking_est	Smoking_se
1	0.03355778	0.1226416
2	-0.4404486	0.2866193
3	0.00751477	0.12542296
4	-0.11468116	0.2152298
5	-0.1234989	0.0924498

Table 2: Cross-sectional estimates for the smoking effect at each of the 5 years, for sample sizes $n = 100$, $n = 500$, and $n = 1000$. The columns `Smoking_est` and `Smoking_se` refer to the estimated coefficient of current-year smoking status and its standard error, respectively. (Note: numbers here are illustrative; actual simulation results may differ.)