BIOS 6312: Modern Biostatistics Methodology II

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 8: Introduction to clustered data

Version: 04/26/2025

Table of Contents

Correlated data

Recap:

- So far in this course, we have emphasized the importance of knowing the assumptions involved in a statistical analysis.
  - Some of the assumptions are imposed by the model (e.g., linearity).
  - Some assumptions are associated with the estimation method (e.g., mean-variance relationship).
- The most frequently invoked assumption thus far has probably been independence of observations (so much the case, that I've largely stopped referencing it).
- Many studies involve observations that cannot reasonably be assumed to be independent.

Correlated data

Some kinds of correlated data:

- Clustered data:
  - Patients within health-care centers.
  - Students within classrooms.
  - Genetic data within patients.
- Longitudinal data:
  - Example: REACH (HbA1c over time).
- Clustered longitudinal data:
  - Patients over time within health-care centers.

Correlated data

Motivation of methods for correlated data:

- There are examples in which correlated data should be viewed as an opportunity to be leveraged:
  - For reasons of clinical relevance (e.g., understanding patient outcome trajectories over time).
  - For statistical advantage (e.g., increasing precision of estimation).
- Other times, correlation may be seen as more of a nuisance, but accounting for it is important nevertheless.
  - Example: it may be more convenient to enroll $n = 50$ patients at each of five study sites rather than $N = 250$ patients at a single site.
  - Example: spatial correlations often play a *huge* role in why political polls/election models can overstate degree of confidence.

Correlated data

Motivation of methods for correlated data:

- There are courses devoted to statistical methods for correlated data.
- I will focus nearly exclusively on one class of methods to handle clustering in linear models for continuous outcomes in these notes.
- Understand that while *some* of the concepts and procedures presented in these notes readily generalize to non-continuous outcomes, many will not.

Table of Contents

Example: REACH

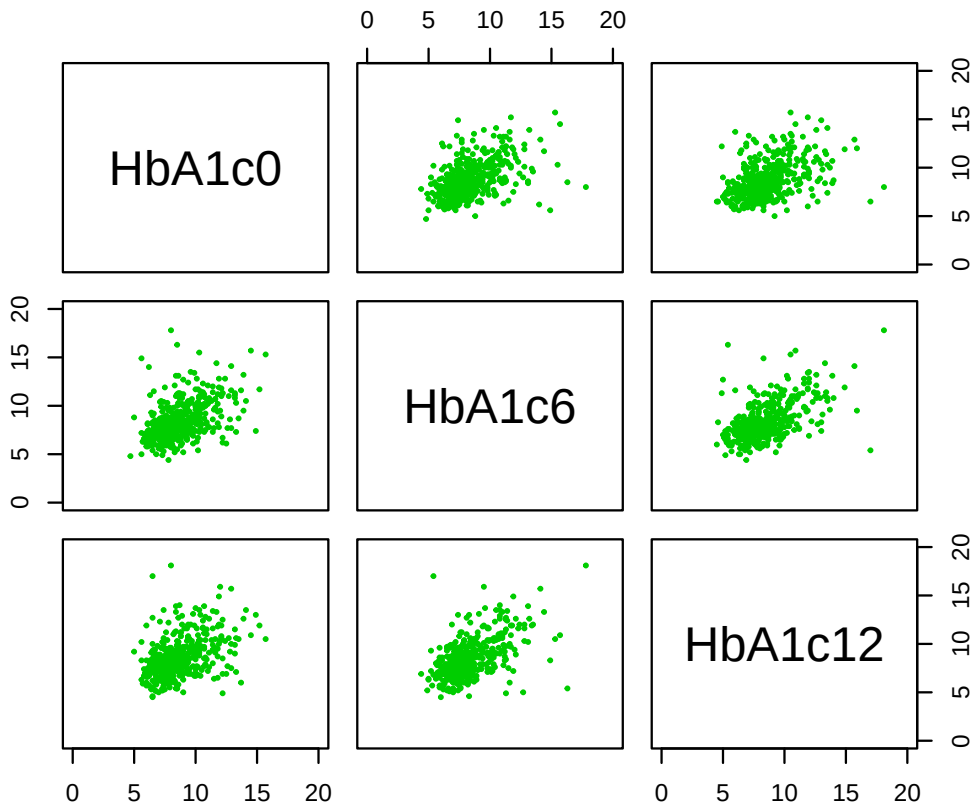
- Recall the REACH study, which we will use as an anchoring example in this set of notes.
- Randomized controlled trial of patients with type 2 diabetes.
  - X: intervention group (0 = control; 1 = REACH).
    - ⋆ Time-stable, deterministic.
  - Y: HbA1c (%).
    - ⋆ Measured six months and twelve months post-intervention.
    - ⋆ Also measured at baseline, but we have largely treated it as an adjustment covariate due to its pre-intervention status.
- Let's explore the REACH data longitudinally.

REACH: Descriptives on HbA1c over time and by group

```
## Read in data set
dat <- read.csv("reach.csv")

## Descriptive statistics
> descrip(cbind(a1c0 = dat$a1c.0,
                a1c6 = dat$a1c.6,
                a1c12 = dat$a1c.12),
          strata = dat$reach)

                    N  Msng    Mean  Std Dev     Min     25%     Mdn     75%     Max
a1c0:  All        505    10    8.62     1.89    4.70    7.20    8.30    9.75    15.7
a1c0:    Str  0   252     3    8.54     1.87    4.70    7.20    8.20    9.50    15.7
a1c0:    Str  1   253     7    8.71     1.91    5.60    7.23    8.40    9.80    15.2
a1c6:  All        505    63    8.40     2.02    4.40    7.00    8.00    9.40    17.8
a1c6:    Str  0   252    31    8.72     2.21    4.80    7.30    8.20    9.80    17.8
a1c6:    Str  1   253    32    8.09     1.76    4.40    6.90    7.80    9.00    14.9
a1c12:  All       505    62    8.57     2.09    4.50    7.10    8.20    9.60    18.1
a1c12:   Str  0   252    22    8.58     2.12    4.50    7.10    8.40    9.67    18.1
a1c12:   Str  1   253    40    8.56     2.06    4.60    7.20    8.20    9.60    17.0
```

REACH: Pairwise scatter plots

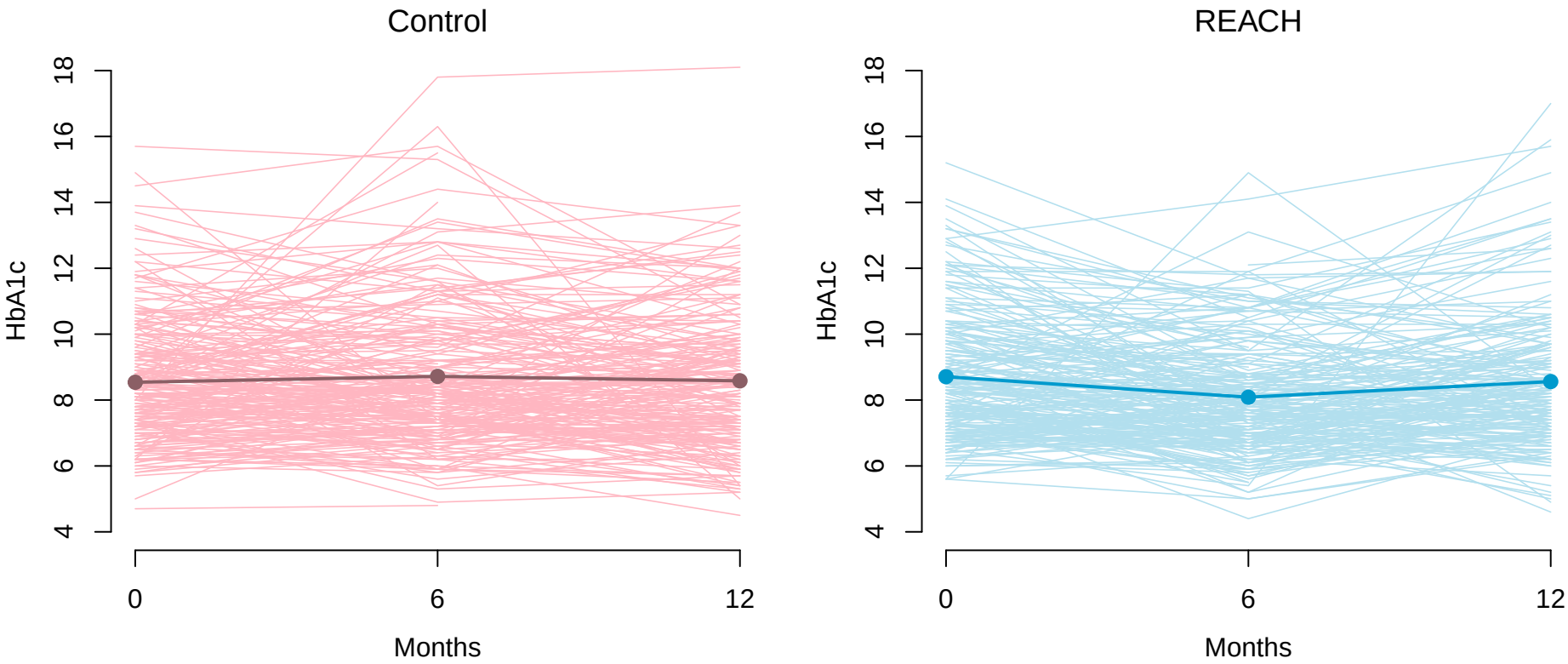
```
## Pairwise scatter plot
pairs(cbind(HbA1c0 = dat$a1c.0, HbA1c6 = dat$a1c.6, HbA1c12 = dat$a1c.12),
      xlim = c(0,20), ylim = c(0,20),
      pch = 20, cex = 0.5, col = "green3")
```



REACH: Spaghetti plots

```r
## Control data (as an example); you can do REACH by analogy
ldat0 <- data.frame(dat$a1c.0[dat$reach == 0], dat$a1c.6[dat$reach == 0],
                    dat$a1c.12[dat$reach == 0])

## Time points
time_points <- c(1, 2, 3)

# Set up empty plot
plot(NA, xlim = c(1, 3), ylim = c(4,18), xlab = "Months", ylab = "HbA1c",
     xaxt = "n", main = expression(paste("Control")),
     frame.plot = FALSE)

## Add axis
axis(1, at = time_points, labels = c("0", "6", "12"))

# Add lines for each subject
apply(ldat0, 1, function(row) {
  lines(time_points, row, col = "lightpink", lwd = 0.8)
})

## Add trajectory of sample means
lines(c(1:3), colMeans(ldat0, na.rm = TRUE), col = "lightpink4",
      lwd = 2, type = "o", pch = 20, cex = 1.5)
```

REACH: Spaghetti plots



Highlights means and heterogeneity of longitudinal trajectories.

REACH: Reshaping the data

```
## Preview wide data (subset of columns)
> head(dat)[,c(1:5, 12:17)]
  id reach fams age gender a1c.0 a1c.6 a1c.12 sdsca.0 sdsca.6 sdsca.12
1  1     0    0  36      1  10.3  15.5     NA     7.0      NA       NA
2  2     0    0  51      0   6.2  14.0     NA     6.0     5.6      6.4
3  3     0    0  48      0   6.8    NA    5.8     4.7     4.2      6.0
4  4     0    0  59      0   7.8    NA    8.0     7.8     6.8      6.3
5  5     0    0  62      0   7.4   5.3    5.7     6.2     5.9      7.8
6  6     0    0  62      0   8.7  11.3    8.4     3.5     4.6      6.7

## Reshape data
dat.long <- reshape(dat,
                    varying = list(c("a1c.0", "a1c.6", "a1c.12"),
                                   c("sdsca.0", "sdsca.6", "sdsca.12")),
                    direction = "long")

## Order to keep those of the same ID together (important for later)
dat.long <- dat.long[order(dat.long$id),]
row.names(dat.long) <- NULL
names(dat.long)[13:14] <- c("a1c", "sdsca")
```

REACH: Viewing in long format

```
## Preview data
> head(dat.long)
  id reach fams age gender raceeth educyears dmdur insulin minority disadv time  a1c sdsca
1  1     0    0  36      1       2         15     1       0        1      0    1 10.3   7.0
2  1     0    0  36      1       2         15     1       0        1      0    2 15.5    NA
3  1     0    0  36      1       2         15     1       0        1      0    3   NA    NA
4  2     0    0  51      0       3         19     5       1        1      1    1  6.2   6.0
5  2     0    0  51      0       3         19     5       1        1      1    2 14.0   5.6
6  2     0    0  51      0       3         19     5       1        1      1    3   NA   6.4
```

Example: REACH

```
## Follow-up data only
dat.follow <- subset(dat.long, dat.long$time != 1)

## Preview data
> head(dat.follow)
  id reach fams age gender raceeth educyears dmdur insulin minority disadv time  a1c sdsca
2  1     0    0  36      1       2        15     1       0        1      0    2 15.5    NA
3  1     0    0  36      1       2        15     1       0        1      0    3   NA    NA
5  2     0    0  51      0       3        19     5       1        1      1    2 14.0   5.6
6  2     0    0  51      0       3        19     5       1        1      1    3   NA   6.4
8  3     0    0  48      0       3        15     4       0        1      1    2   NA   4.2
9  3     0    0  48      0       3        15     4       0        1      1    3  5.8   6.0
```

Example: REACH

- Variables:
  - X: intervention group (0 = control; 1 = REACH).
  - Y: HbA1c (%).
- Our prior analyses of REACH have only made use of a single value of HbA1c as an outcome at a time.
- Therein lies the first simple approach to addressing repeated measures that is still technically correct, which is to break the longitudinal problem down into a sequence of estimation problems that do not involve correlated outcomes.
- This approach is valid for estimating treatment effects over time, but fundamentally limited in that it will not allow you to . . .
  - . . . leverage efficiency gains from correlated outcomes (with the exception that you can include *baseline* HbA1c as a covariate).
  - . . . perform comparisons of mean outcomes/effects at different times.

Example: REACH

- Related approaches: aggregate/collapse measures that are repeated over time.
  1. Change from baseline:
     - ⋆ Potential for efficiency gains as compared to an analysis of the raw outcome if there is a (positive) correlation.
     - ⋆ However, there is no benefit to effect estimation if you've already adjusted for baseline.
  2. Average measures within subject:
     - ⋆ May make sense when you expect the outcomes to have the same mean (e.g., two runs of an antibody assay on the same person taken from a common blood sample).
     - ⋆ This approach does not make sense for REACH, in which the source of the clustering is longitudinal measurement.

Table of Contents

GEE for continuous outcomes

Example: REACH

- Let's focus first on writing down a model that encodes the longitudinal nature of the data.
- Notation needs to accommodate repeated measures:
  - X: intervention group (0 = control; 1 = REACH).
  - Y_t: HbA1c (%) at time t, where $t = 0, 1, 2$.
- Here t indexes time (Y_0 denotes baseline HbA1c; Y_1 and Y_2 denote the follow-up HbA1c outcomes). Note that X does not need a subscript t in this example because the randomization doesn't change from baseline.
- Consider the following saturated model ($t = 1, 2$):

$$\mathrm{E}[Y_t|X = x] = \beta_0 + \beta_1 x + \beta_2 1(t = 2) + \beta_3 1(t = 2)x.$$

- We can interpret each of its coefficients.

GEE for continuous outcomes

Example: REACH

- Variables:
  - X: intervention group ($0 =$ control; $1 =$ REACH).
  - Y_t: HbA1c (%) at time t, where $t = 0, 1, 2$.
- Model: $\mathrm{E}[Y_t|X = x] = \beta_0 + \beta_1 x + \beta_2 1(t = 2) + \beta_3 1(t = 2)x$, $t = 1, 2$.
- Interpretation of coefficients:
  - β_0: $\mathrm{E}[Y_1|X = 0]$
    - ⋆ Mean HbA1c for control stratum six months post-baseline.
  - β_1: $\mathrm{E}[Y_1|X = 1] - \mathrm{E}[Y_1|X = 0]$
    - ⋆ Effect of REACH on mean HbA1c six months post-baseline.
  - β_2: $\mathrm{E}[Y_2|X = 0] - \mathrm{E}[Y_1|X = 0]$
    - ⋆ Mean change in HbA1c from six two twelve months post-baseline within the control stratum.
  - β_3: $(\mathrm{E}[Y_2|X = 1] - \mathrm{E}[Y_2|X = 0]) - (\mathrm{E}[Y_1|X = 1] - \mathrm{E}[Y_1|X = 0])$
    - ⋆ Change in effect of REACH on mean HbA1c from six to twelve months post-baseline.

GEE for continuous outcomes

Example: REACH

- Variables:
  - X: intervention group ($0 =$ control; $1 =$ REACH).
  - Y_t: HbA1c (%) at time t, where $t = 1, 2$.
- Model: $\mathrm{E}[Y_t|X = x] = \beta_0 + \beta_1 x + \beta_2 1(t = 2) + \beta_3 1(t = 2)x$
- We will discuss *one* approach to estimation of $\boldsymbol{\beta}$ known as generalized estimating equations (GEE).

GEE FOR CONTINUOUS OUTCOMES

Generalized estimating equations (GEE): The basic idea

- Let $i = 1, \ldots, n$ denote independently sampled clusters, each with T_i observations.
- Focus is on the mean model (expressed in two ways below):

$$\begin{aligned} \mathrm{E}[Y_{it}|\mathbf{X}_{it} = \mathbf{x}_{it}] &= \mathbf{x}_{it}^T\boldsymbol{\beta} \\ \mathrm{E}[\mathbf{y}_i|\mathbf{X}_i] = \boldsymbol{\mu}_i(\boldsymbol{\beta}) &= \mathbf{X}_i\boldsymbol{\beta}. \end{aligned}$$

- The basic idea is to construct a set of equations that will allow us to solve for $\boldsymbol{\beta}$, and then use a variance formula that accounts for repeated measures on subjects.
- The OLS equations that we already know actually work just fine to accomplish the former goal; it's just that the variance formula will then need to be updated to reflect the correlated data.

GEE with working independence: Estimation

- Mean model: $\mathrm{E}[\mathbf{y}_i|\mathbf{X}_i] = \mathbf{X}_i\boldsymbol{\beta}$.
- **Step 1**: Solve the following equations for $\boldsymbol{\beta}$:

$$\sum_{i=1}^{n} \mathbf{X}_i^T(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \quad = \quad \mathbf{0} \Rightarrow \widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{X}_i^T\mathbf{y}_i\right).$$

- Notes about Step 1:
  - This is exactly the same as the ordinary least squares estimate, and treats the observations as if they were all independent.
  - Note that the number of observations may vary across independent clusters.

GEE with working independence: Estimation

- Mean model: $\mathrm{E}[\mathbf{y}_i|\mathbf{X}_i] = \mathbf{X}_i\boldsymbol{\beta}$.
- **Step 2**: Use a sandwich variance that reflects the independence of subjects (but not observations within subjects). Letting $\widehat{\boldsymbol{\epsilon}}_i = \mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}$,

$$\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}] = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^T\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{\epsilon}}_i^T\mathbf{X}_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i\right)^{-1}$$

- Notes about Step 2:
  - This sandwich variance allows each independent subject to contribute to estimation of the covariance matrix but makes no presumption that observations within a subject are independent.
  - This is sometimes called the "cluster robust" variance.

GEE with working independence: Estimation

- The previous slides outline a procedure known as GEE with a "working independence correlation structure."
- The idea is that we can treat all observations across all subjects as independent to start with (with the understanding that the resulting estimator is consistent), and then create corresponding standard errors that are valid.
- It is indeed possible to start with *alternative* working correlation structures (and even a working mean-variance relationship).

GEE: Covariance structure

- $\mathbf{R} = \mathbf{R}(\boldsymbol{\alpha})$: Correlation structure between time points (note that $\boldsymbol{\alpha}$ is not typically a parameter of interest and $\mathbf{R}_{t,t} = 1$ for all t).
  - Independence: $\mathbf{R}_{t,t'} = \text{Corr}(Y_{it}, Y_{it'}|\mathbf{X}_i = \mathbf{x}_i) = 0$ ($\mathbf{R} = \mathbf{I}$).
  - Exchangeable: $\mathbf{R}_{t,t'} = \text{Corr}(Y_{it}, Y_{it'}|\mathbf{X}_i = \mathbf{x}_i) = \alpha$.
    - ⋆ Example: $\begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix}$.
  - Auto-regressive: $\mathbf{R}_{t,t'} = \text{Corr}(Y_{it}, Y_{it'}|\mathbf{X}_i = \mathbf{x}_i) = \alpha^{|t-t'|}$.
    - ⋆ Example: $\begin{pmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{pmatrix}$.
  - Unstructured: $\mathbf{R}_{t,t'} = \text{Corr}(Y_{it}, Y_{it'}|\mathbf{X}_i = \mathbf{x}_i) = \alpha_{t,t'}$.
    - ⋆ Example: $\begin{pmatrix} 1 & \alpha_1 & \alpha_2 \\ \alpha_1 & 1 & \alpha_3 \\ \alpha_2 & \alpha_3 & 1 \end{pmatrix}$.

GEE for continuous outcomes

GEE: Working covariance (as a new take on an old concept)

- GEE with working independence can be described as using the ordinary least squares estimator for independent outcomes and following up with a sandwich variance suitable for repeated outcomes.
- Other forms of GEE: propose a correlation structure and tweak the cluster robust variance accordingly.
  - Why might we want to consider alternatives to an approach based on working independence?

GEE for continuous outcomes

GEE: Working covariance

- The three most common flavors for correlation structure:
  - Independence.
  - Exchangeable.
  - Auto-regressive.
- Among these, the working independence model is the most frequently used. Although not generally efficient, working independence has the upper hand in number of settings when dealing with non-continuous outcomes—specifically with respect to time-dependent covariates and assumptions surrounding missing data.

GEE: Covariance structure

- $\mathbf{R} = \mathbf{R}(\boldsymbol{\alpha})$: Correlation structure between time points (note that $\boldsymbol{\alpha}$ is not typically a parameter of interest and $\mathbf{R}_{t,t} = 1$ for all t).
  - Independence: $\mathbf{R}_{t,t'} = \text{Corr}(Y_{it}, Y_{it'}|\mathbf{X}_i = \mathbf{x}_i) = 0$ ($\mathbf{R} = \mathbf{I}$).
  - Exchangeable: $\mathbf{R}_{t,t'} = \text{Corr}(Y_{it}, Y_{it'}|\mathbf{X}_i = \mathbf{x}_i) = \alpha$.
    - ⋆ Example: $\begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix}$.
  - Auto-regressive: $\mathbf{R}_{t,t'} = \text{Corr}(Y_{it}, Y_{it'}|\mathbf{X}_i = \mathbf{x}_i) = \alpha^{|t-t'|}$.
    - ⋆ Example: $\begin{pmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{pmatrix}$.
  - Unstructured: $\mathbf{R}_{t,t'} = \text{Corr}(Y_{it}, Y_{it'}|\mathbf{X}_i = \mathbf{x}_i) = \alpha_{t,t'}$.
    - ⋆ Example: $\begin{pmatrix} 1 & \alpha_1 & \alpha_2 \\ \alpha_1 & 1 & \alpha_3 \\ \alpha_2 & \alpha_3 & 1 \end{pmatrix}$.

GEE for continuous outcomes

GEE: R

- In R, we'll use the `geese()` function in the `geepack` library.
- Dealing with stratum-specific means and effects: We can use my customized `LinCom()` function.
- Dealing with omnibus tests: We can use my customized `JointTest()` function.

Recall: Custom `LinCom()` function

```r
## Linear combination of parameters to test (eform optional)
LinCom <- function(idxtest, mults, coefs, varmat, eform = FALSE) {
  R <- matrix(0, nrow = 1, ncol = length(coefs))
  for (q in 1:length(idxtest)) {R[1,idxtest[q]] <- mults[q]}
  wld <- as.numeric(t(R %*% coefs) %*% solve(R %*% varmat %*% t(R)) %*% (R %*% coefs))
  pval <- 1 - pchisq(wld, df = 1)
  Est <- R %*% coefs
  CI.Lo <- R %*% coefs - qnorm(0.975)*sqrt(R %*% varmat %*% t(R))
  CI.Hi <- R %*% coefs + qnorm(0.975)*sqrt(R %*% varmat %*% t(R))
  if (eform == TRUE){
    return(c(EST = exp(Est), CI.LO = exp(CI.Lo), CI.HI = exp(CI.Hi), P = pval))
  }
  if (eform == FALSE){
    return(c(EST = Est, CI.LO = CI.Lo, CI.HI = CI.Hi, P = pval))
  }
}
```

Recall: Custom `JointTest()` function

```r
## Omnibus test of multiple parameters
JointTest <- function(idxtest, coefs, varmat) {
  R <- matrix(0, nrow = length(idxtest), ncol = length(coefs))
  for (q in 1:length(idxtest)) {R[q,idxtest[q]] <- 1}
  wld <- as.numeric(t(R %*% coefs) %*% solve(R %*% varmat %*% t(R)) %*% (R %*% coefs))
  pval <- 1 - pchisq(wld, df = length(idxtest))
  return(c(chi2.stat = wld, p = pval))
}
```

Example: REACH

- Variables:
  - X: intervention group ($0 =$ control; $1 =$ REACH).
  - Y_t: HbA1c (%) at time t, where $t = 1, 2$.
- Model: $\mathrm{E}[Y_t|X = x] = \beta_0 + \beta_1 x + \beta_2 1(t = 2) + \beta_3 1(t = 2)x$.
- Let's use GEE with working independence to estimate $\boldsymbol{\beta}$. Specifically, let's focus on estimating the six-month and twelve-month effects.
  - Six-month effect: β_1.
  - Twelve-month effect: $\beta_1 + \beta_3$.
  - We can also jointly test for any effect (may or may not be wise).
  - We can also compare effects between times via β_3.

Example: REACH

```
## Time is a factor variable (not to be treated continuously here)
dat.follow$time <- factor(dat.follow$time)

## Fit model (abridged output)
model <- geese(a1c ~ time*reach, id = id, corstr = "indep", data = dat.follow)
> model

Call:
geese(formula = a1c ~ time * reach, id = id, data = dat.follow,
    corstr = "indep")

Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian

 Coefficients:
(Intercept)       time3       reach time3:reach
     8.7181     -0.1346     -0.6330      0.6106
```

REACH: Six-month effect

```
## Test of interest (six-month effect)
> LinCom(idxtest = c(3), mults = c(1), coefs = model$beta,
         varmat = model$vbeta)

       EST      CI.LO      CI.HI           P
-0.6330317 -1.0047386 -0.2613247  0.0008442
```

REACH: Twelve-month effect

```
## Test of interest (twelve-month effect)
> LinCom(idxtest = c(3,4), mults = c(1,1),
         coefs = model$beta, varmat = model$vbeta)

     EST    CI.LO    CI.HI         P
-0.02245 -0.41048  0.36559  0.90974
```

REACH: Test of any effect

```
## Test of interest
> JointTest(idxtest = c(3,4), coefs = model$beta,
            varmat = model$vbeta)

chi2.stat          p
1.424e+01 8.095e-04
```

REACH: Test of change in effect over time

```
> LinCom(idxtest = c(4), mults = c(1), coefs = model$beta,
         varmat = model$vbeta)

      EST    CI.LO    CI.HI        P
0.610586 0.227700 0.993472 0.001775
```

GEE for continuous outcomes

REACH: Further discussion

- Note that you could have recovered the same estimates from fitting separate linear models.
- Part of the benefit of having everything in one unified model is that you can compare effects over time.
- Another benefit is that you can *adjust* for baseline covariates in a way that controls degrees of freedom.

Additional thoughts:

- Reminder: *Some*, but not *all* of the procedures generalize to non-continuous outcomes.
- Keep in mind the things that we did *not* discuss:
  - Mixed effects models.
  - Non-continuous outcomes.
  - More sophisticated correlation structures.
- Course on longitudinal data analysis will also typically also feature in-depth explorations of:
  - Time-dependent covariates.
  - Nonlinearity, non-binary outcomes, and non-collapsibility.
  - Assumptions surrounding missingness.
- There is a reason why correlated data methods typically get an entire semester of treatment.

Summary

This unit:

- Data visualization and reshaping.
- Longitudinal data methods as an opportunity to estimate/compare quantities that cannot be compared in its absence.
- Generalized estimating equations with working independence.

Summary

So far:

- Review.
- Simple linear regression.
- Multiple linear regression (foundations).
- Multiple linear regression (interactions and strata).
- Transformations and basis expansions.
- Regression with binary outcomes.
- Regression with nominal, ordinal, and count outcomes.
- Introduction to clustered data.

Summary

Coming up:

- Methods for time-to-event outcomes.
- Predictive capacity of regression models.