Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Spring 2020
Homework #7 Key

1. Load the data set `reach.csv`. We will consider several longitudinal analyses of A1c.

   (a) Report point estimates and 95% confidence intervals for the effect of REACH at six and twelve months based on two separate regression models, Models (1) and (2):

   $$\mathbf{E}[\texttt{A1c}_6|\texttt{REACH}, \texttt{A1c}_0] = \beta_0 + \beta_1\texttt{REACH} + \beta_2\texttt{A1c}_0, \text{ and} \quad (1)$$
   $$\mathbf{E}[\texttt{A1c}_{12}|\texttt{REACH}, \texttt{A1c}_0] = \beta_0 + \beta_1\texttt{REACH} + \beta_2\texttt{A1c}_0 \quad (2)$$

   Ans: We fit these models using OLS linear regression with robust standard errors to allow heteroscedasticity. We characterize the treatment effect in terms of how much lower the mean A1c is at a particular time point for the REACH group, relative to control. At six months, we estimate the baseline-A1c-adjusted treatment effect to be 0.716% (95% CI: [0.388%, 1.04%]; $p < 0.001$). At twelve months, we estimate the baseline-A1c-adjusted treatment effect to be 0.0860% (95% CI: [-0.267%, 0.439%]; $p = 0.633$).

   (b) Consider the following model with A1c outcomes at times $t = 6$ and $t = 12$ months:

   $$\mathbf{E}[\texttt{A1c}_t|\texttt{REACH}, \texttt{A1c}_0] = \beta_0 + \beta_1\texttt{REACH} + \beta_2\texttt{A1c}_0. \quad (3)$$

   Provide a plain-language, but specific interpretation for β_1 from Model (3). Then, use GEE with a working exchangeable correlation structure to estimate β_1, and hence perform an analysis to evaluate the overall effect of REACH. Considering results from part (a), explain briefly why Model (3) is very seriously flawed. *Hint*: what major assumptions are imposed by Model (3) but *not* by separate consideration of Models (1) and (2)?

   Ans: β_1 corresponds to the difference in mean follow-up A1c between the subgroups of the same baseline A1c but differing in whether they received REACH or control. Based on a GEE fit using a working exchangeable correlation structure, we obtain an estimate of the treatment effect to be 0.438% (95% CI: [0.157%, 0.718%]; $p = 0.002$). The reason this is severely flawed is that this treatment effect is assumed the same at both time points, when our results from part (a) very clearly suggest that this is not the case. Less serious of a flaw in this model is the assumption that the association between baseline A1c and mean follow-up A1c does not depend upon the time of the outcome.

   (c) Consider instead the following model, again with outcomes at times $t = 6$ and $t = 12$ months:

   $$\mathbf{E}[\texttt{A1c}_t|\texttt{REACH}, \texttt{A1c}_0] = \beta_0 + \beta_1\texttt{REACH} + \beta_2\texttt{1(t = 12)} + \beta_3\texttt{1(t = 12)} \times \texttt{REACH} + \beta_4\texttt{A1c}_0. \quad (4)$$

   What major assumptions are imposed by Model (4) but *not* by Models (1) and (2)?

   Ans: This model presumes that the association between baseline A1c and mean follow-up A1c does not depend upon the time of the outcome.

(d) State a hypothesis in terms of the coefficients of Model (4) that could be used to evaluate whether there is sufficient evidence of an overall REACH effect. Use GEE with a working exchangeable correlation structure to test that hypothesis.

Ans: $H_0 : \beta_1 = \beta_3 = 0$. Based on a test of this hypothesis, we obtain sufficient evidence to reject the null hypothesis of no overall treatment effect ($p < 0.001$).

(e) In plain language, characterize the hypothesis test $H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$. Based on the GEE model you fit in part (d), test that hypothesis.

Ans: This is a test of whether the baseline-A1c-adjusted effect of REACH on mean six-month A1c differs from that at twelve-months. Based on the GEE model that we've fit, we obtain evidence to reject the null hypothesis of equal treatment effects, further slamming the door on the reasonableness of Model (3) ($p = 0.001$). Importantly, such strength in my wording is warranted by the the clinical relevance of the difference in the treatment effects.

(f) State a hypothesis in terms of the coefficients of Model (4) that could be used to evaluate whether there is sufficient evidence of an effect of REACH at six months. Based on the GEE model you fit in part (d), test that hypothesis.

Ans: $H_0 : \beta_1 = 0$. There is sufficient evidence of such a treatment effect ($p < 0.001$). Note the point estimate, 0.750%, and the 95% confidence interval, [0.426%, 1.07%], and how similar they are to those in part (a), the latter having a slightly narrower confidence interval.

(g) Propose a simple alteration to Model (4) that could be used to evaluate whether there is sufficient evidence of an effect of REACH at *twelve* months; perform this hypothesis test based on analogous GEE-based estimation of the modified model.

Ans: This idea should be somewhat familiar. Instead, reverse the role of six- and twelve-months by defining an indicator variable for six-months and fit the analogous model. The treatment effect was estimated to be 0.116% (95% CI: [-0.234%, 0.467%]; $p = 0.515$); note again how this compared to part (a). There is not sufficient evidence of a twelve-month effect.

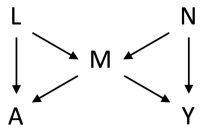
(h) Do the typical concerns associated with using a non-independence working correlation structure apply in this example? Very briefly explain.

Ans: You could note that we're using the identify link or note that the variables are time-stable (REACH and baseline A1c) or deterministic (time). Either way, we are not concerned.

(i) Repeat part (d), instead using a random intercepts model instead of GEE. Briefly highlight differences (if any) in coefficient interpretation as compared to the GEE model of part (d).

Ans: $H_0 : \beta_1 = \beta_3 = 0$. We obtain sufficient evidence to reject the null hypothesis of no overall treatment effect ($p < 0.001$). This could be interpreted marginally (population-averaged) or conditionally (within-subject) in light of the fact that there is no confounding (RCT).

2. *Optional problem*: This optional problem is purely pedagogical. Suppose you have measured the following: A (indicator of low education), diabetes status (Y), childhood family income (L), mother's diabetes status (M), and mother's genetic risk for diabetes (N). The following causal graph is proposed:



(a) Which backdoor paths from Y into A, if any, are blocked?

Ans: The path $Y \longleftarrow N \longrightarrow M \longleftarrow L \longrightarrow A$ is blocked by the collider, M.

(b) Are any unblocked backdoor paths created by conditioning on M alone?

Ans: Yes. A path is opened between L and N.

(c) Are A and Y d-separated conditional on L and M?

Ans: Yes. All paths are blocked after having conditioned on L and M.

(d) Are A and Y conditionally independent given M and N?

Ans: Yes, because A and Y are d-separated given M and N.

(e) Which of the following statements is/are true? (I) $Y \perp\!\!\!\perp A|M, N$; (II) $Y^a \perp\!\!\!\perp A|M, N$.

Ans: Both are true. The former is true because of the lack of a direct arrow from $A \longrightarrow Y$.

(f) In the context of this problem:

I. State the interpretation of $\Delta = \mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$ and its value based on the DAG.

Ans: $\Delta = 0$, the population average causal effect of A on Y.

II. Would you expect $\widehat{\mathbf{E}}[Y|A = 1] - \widehat{\mathbf{E}}[Y|A = 0]$ to be unbiased for Δ? Explain.

Ans: No, because not all backdoor paths from Y are blocked.

(g) Place a variable U_1 on the DAG such that, when unaccounted for, Δ is still identifiable.

Ans: $U_1 \longrightarrow Y$.

(h) Place a variable U_2 on the DAG such that, when unaccounted for, Δ is *not* identifiable.

Ans: $A \longleftarrow U_2 \longrightarrow Y$.