The Baseline Observation
BIOS 667

Bahjat F. Qaqish
Department of Biostatistics
CB 7420, McGavran Greenberg Hall
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420
email: qaqish@bios.unc.edu
www.bios.unc.edu/~qaqish

The Question

- A 2-group study, $x_i = 0$ versus $x_i = 1$
- Randomized:
  Control or placebo or group 0: $x_i = 0$
  Active treatment or group 1: $x_i = 1$
- Non-randomized (Observational):
  Drug A: $x_i = 0$
  Drug B: $x_i = 1$
- Non-randomized (Observational):
  Non-smoker: $x_i = 0$
  Smoker: $x_i = 1$
- Baseline response: Y_{i0} at time 0
  Follow up response: Y_{i1} at time 1

- Since x_i is either 0 or 1, we can write
  $\mathrm{E}[Y_{i1}|x_i] = \delta_1 + \delta_2 x_i$
  $\mathrm{E}[Y_{i1}|x_i = 0] = \delta_1$
  $\mathrm{E}[Y_{i1}|x_i = 1] = \delta_1 + \delta_2$

- The treatmet contrast (treatment effect) is
  $\delta_2 = \mathrm{E}[Y_{i1}|x_i = 1] - \mathrm{E}[Y_{i1}|x_i = 0]$

- The difference between the group means at time 1:
  $\delta_2 = \mathrm{E}[Y_{i1}|x_i = 1] - \mathrm{E}[Y_{i1}|x_i = 0]$

- δ_2 is the *marginal* treatmet contrast or difference between group means at time 1
- δ_2 is the key parameter of interest
- δ_2 can be estimated by regressing Y_{i1} on x_i, i.e. fitting $\mathrm{E}[Y_{i1}|x_i] = \delta_1 + \delta_2 x_i$

- Question:
  Adjust for baseline by regressing Y_{i1} on x_i and Y_{i0}???
- Would the slope for x_i still estimate δ_2???

The Answers

- Start by finding $\mathrm{E}[Y_{i1}|Y_{i0}, x_i]$
- Assume bivariate normality of (Y_{i0}, Y_{i1}); i.e. one bivariate normal for $x_i = 0$ and another for $x_i = 1$
- $\mathrm{E}[Y_{i1}|Y_{i0}, x_i = 0] = \alpha_1 + \alpha_2 Y_{i0}$ follows from bivariate normality
- $\mathrm{E}[Y_{i1}|Y_{i0}, x_i = 1] = \beta_1 + \beta_2 Y_{i0}$ follows from bivariate normality
- Combine into a single equation:

$$\mathrm{E}[Y_{i1}|Y_{i0}, x_i] \;=\; (\alpha_1 + \alpha_2 Y_{i0}) + \{(\beta_1 - \alpha_1) + (\beta_2 - \alpha_2)Y_{i0}\}x_i$$

- $\mathrm{E}[Y_{i1}|Y_{i0}, x_i]$ contains an interaction term, $x_i Y_{i0}$
- The *conditional* treatmet contrast

$$(\beta_1 - \alpha_1) + (\beta_2 - \alpha_2)Y_{i0}$$

  depends on Y_{i0}
- Answer (part 1): Regressing Y_{i1} on x_i and Y_{i0} would be correct if $\beta_2 - \alpha_2 = 0$, i.e. if there is "no interaction"
- Interaction term: $(\beta_2 - \alpha_2)x_i Y_{i0}$
- Make the assumption NI: $\beta_2 - \alpha_2 = 0$
- Now, with NI, we have:

$$\mathrm{E}[Y_{i1}|Y_{i0}, x_i] \;=\; (\alpha_1 + \alpha_2 Y_{i0}) + (\beta_1 - \alpha_1)x_i$$

- Under NI, the *conditional* treatmet contrast is $\beta_1 - \alpha_1$.
- Under NI, if we regress Y_{i1} on x_i and Y_{i0}, the slope for x_i will estimate $\beta_1 - \alpha_1$
- The question now:
  Assuming NI, is $\beta_1 - \alpha_1 = \delta_2$???
- If the answer is yes, then regressing Y_{i1} on x_i and Y_{i0} produces the right answer, the slope for x_i will estimate δ_2
- So, is $\beta_1 - \alpha_1 = \delta_2$???
- To obtain the marginal expectation, apply double-expectation to

$$\mathrm{E}[Y_{i1}|Y_{i0}, x_i] \;=\; (\alpha_1 + \alpha_2 Y_{i0}) + (\beta_1 - \alpha_1)x_i$$

- Need $\mathrm{E}[Y_{i0}|x_i]$
- Since x_i is either 0 or 1, we can write
  $\mathrm{E}[Y_{i0}|x_i] = \gamma_1 + \gamma_2 x_i$

- By double-expectation

$$\begin{aligned}
\mathrm{E}[Y_{i1}|x_i] &= \mathrm{E}[\mathrm{E}[Y_{i1}|Y_{i0}, x_i]] \\
&= \mathrm{E}[\alpha_1 + (\beta_1 - \alpha_1)x_i + \alpha_2 Y_{i0}|x_i] \\
&= \alpha_1 + (\beta_1 - \alpha_1)x_i + \alpha_2 \mathrm{E}[Y_{i0}|x_i] \\
&= \alpha_1 + (\beta_1 - \alpha_1)x_i + \alpha_2(\gamma_1 + \gamma_2 x_i) \\
&= (\alpha_1 + \alpha_2\gamma_1) + (\beta_1 - \alpha_1 + \alpha_2\gamma_2)x_i
\end{aligned}$$

- Recall: $\mathrm{E}[Y_{i1}|x_i] = \delta_1 + \delta_2 x_i$

- $\delta_1 = \alpha_1 + \alpha_2\gamma_1$,
  $\delta_2 = \beta_1 - \alpha_1 + \alpha_2\gamma_2$ or $\beta_1 - \alpha_1 = \delta_2 - \alpha_2\gamma_2$

- Answer (part 2): When regressing Y_{i1} on x_i and Y_{i0}, the slope for x_i will estimate $\beta_1 - \alpha_1 \neq \delta_2$. The bias is $-\alpha_2\gamma_2$.

- Recall:
  α_2 is the slope in the regression of Y_{i1} on Y_{i0} (the same in both groups by NI, $\alpha_2 = \beta_2$).
  γ_2 is the difference between group means at baseline.

  e.g. Positive bias if α_2 and γ_2 have different signs.

- But can that bias be 0?
  $\beta_1 - \alpha_1 = \delta_2$ if $\alpha_2\gamma_2 = 0$

- $\alpha_2\gamma_2 = 0$ if $\alpha_2 = 0$ or $\gamma_2 = 0$ (or both)

- $\alpha_2 = 0$?
  Under NI, this also means $\alpha_2 = \beta_2 = 0$, Y_{i0} is independent of Y_{i1}.
  Not realistic for longitudinal outcomes.

- $\gamma_2 = 0$? This means that $\mathrm{E}[Y_{i0}|x_i]$ does not depend on x_i. This is valid in a randomized study, but not in an observational (non-randomized) study.

Gain?

- If the study is randomized and there is no interaction (parallel regression lines for Y_{i1} on Y_{i0} in the two groups), what do we gain by adjusting for the baseline (regressing Y_{i1} on x_i and Y_{i0}) versus regressing Y_{i1} on x_i only?
- Without going into derivations, the answer is increased precision, smaller variance of $\hat{\delta}_2$.
- In the regression of Y_{i1} on x_i, the "variance" σ^2 is $\text{var}(Y_{i1})$, the marginal variance.
- In the regression of Y_{i1} on x_i and Y_{i0}, the "variance" σ^2 is $\text{var}(Y_{i1}|Y_{i0})$, the conditional variance.
- The main reason for the increased precision is the mathematical fact: conditional variance $\leq$ the marginal variance in the bivariate normal.
- Note: In distributions other than the bivariate normal, the conditional variance can exceed the marginal variance, depending on the specific value of the conditioning variable. Example: If (Y_{i0}, Y_{i1}) is bivariate Bernoulli, then $\text{var}(Y_{i1}|Y_{i0} = t)$ can exceed $\text{var}(Y_{i1})$. Exercise: Develop a numerical example.
- How much do we gain?
- Example: $n - 1$ observations post-baseline, averaged; common variance σ^2, common correlation ρ. Relative efficiency is

$$\frac{1 + (n-1)\rho}{n}$$

  e.g. $n = 4, \rho = 1/3$, relative efficiency $= 0.5$.

Loss?

- Even if the NI assumption is satisfied, in a non-randomized study the slope for x_i, when adjusting for baseline, estimates $\delta_2 - \alpha_2\gamma_2$, while the target quantity is δ_2. BIAS.
- What if the NI assumption is not satisfied in a randomized study? The derivations become more complicated, but the answer is simple: the slope estimate for x_i will be a biased estimator of δ_2.
- No pain, no gain?!