Covariate adjustment and estimation of mean outcome in randomised trials

Jonathan Bartlett
www.thestatsgeek.com

Department of Mathematical Sciences
University of Bath
UK

International Biometric Conference, 13th July 2018

Acknowledgements

- This work was conducted while I was employed at the Statistical Innovation Group at AstraZeneca
- Thanks to Prof. Stijn Vansteelandt, University of Ghent, for helpful input on this work
- Thanks to AstraZeneca Sirocco trial study team for use of data in illustrative example used in the paper

Outline

Motivation

Baseline adjusted mean estimation

Simulations

Conclusions

Motivation

- Consider a randomised trial which samples n patients from a population and randomises patients to control $Z = 0$ or active treatment $Z = 1$
- We measure an outcome Y on each patient, and typically also some baseline covariates X
- Often the primary analysis adjusts for X in the analysis of outcome Y
- This is typically performed by fitting a regression model for Y, with X and Z as covariates
- By using X, we can adjust our treatment effect estimate for chance imbalances in the X distribution between randomised groups, thereby improving statistical power

Motivation

- As well as the estimated treatment effect, the crude mean outcome in each treatment group is also almost always (and should be) reported. For treatment group $Z = z$, this is:

$$\hat{\mu}_1(z) = \frac{\sum_{i=1}^{n} 1(Z_i = z) Y_i}{\sum_{i=1}^{n} 1(Z_i = z)}$$

- Because of randomisation, $\hat{\mu}_1(z)$ unbiasedly estimates

$$E(Y|Z = z) = E(Y^z) := \mu(z)$$

  where Y^z is a patient's potential outcome were they to receive treatment z

- $\mu(z)$ is the average outcome were the population all to be assigned to receive treatment z

Questions

- Can we use our covariate adjusted model to estimate $\mu(z)$?
- If so, how, and what would be the benefit of doing so?

Outline

Motivation

Baseline adjusted mean estimation

Simulations

Conclusions

Outcome regression model

- Assume an outcome model is defined, part of which specifies that (for example)

  $$g(E(Y|X,Z)) = \beta_0 + \beta_X^T X + \beta_Z Z$$

  for some link function $g(.)$
- We fit the model to the trial data, and obtain estimates $\hat{\beta}_0$, $\hat{\beta}_X$, $\hat{\beta}_Z$

Mean outcome at the mean covariate values

- An approach sometimes used to obtain X adjusted estimates of mean outcome under treatment z is to calculate

$$g^{-1}(\hat{\beta}_0 + \hat{\beta}_X^T \hat{\mu}_X + \hat{\beta}_Z z),$$

  setting X equal to its sample mean $\hat{\mu}_X$
- Assuming the outcome model is correctly specified, this is a consistent estimator of $E(Y|X = \mu_X, Z = z)$
- But in general, even though X and Z are independent in a randomised trial,
  $\mu(z) = E(Y^z) = E(Y|Z = z) \neq E(Y|X = \mu_X, Z = z)$
- The quantity being targeted arguably makes little sense for non-linear models when some covariates are categorical

Baseline adjusted estimates of $\mu(z)$

- To estimate $\mu(z)$ using our model which adjusts for X, we note that

$$\mu(z) = E(Y^z) = E\{E(Y^z|X)\}$$

- The inner expectation can be estimated using our fitted regression model by $g^{-1}(\hat{\beta}_0 + \hat{\beta}_X X + \hat{\beta}_Z z)$
- The outer expectation is then with respect to the distribution of X
- This motivates the estimator

$$\hat{\mu}_2(z) = \frac{1}{n} \sum_{i=1}^{n} g^{-1}(\hat{\beta}_0 + \hat{\beta}_X X_i + \hat{\beta}_Z z)$$

Qu and Luo 2015

- $\hat{\mu}_2(z)$ was proposed in the randomised trials context in 2015 by Qu and Luo
- They proposed it as an estimator of the following parameter:

$$n^{-1} \sum_{i=1}^{n} g^{-1}(\beta_0 + \beta_X X_i + \beta_Z z)$$

- This 'parameter' is defined in terms of the covariate values of the particular sample of patients and the population parameters $\beta_0, \beta_X, \beta_Z$. Its value thus varies from trial sample to trial sample, if the X_i are not fixed
- Its values also differs, even if the outcome model is correctly specified, from $n^{-1} \sum_{i=1}^{n} Y^z$, the mean outcome for the trial sample were they to be given treatment z
- We will instead focus on estimation of the population parameter $\mu(z)$...

Intuition for $\hat{\mu}_2(z)$

- $\mu_2(z)$ uses predictions from all patients to estimate $\mu(z)$, and not only those randomised to $Z = z$
- We have the potential to gain a more precise estimate of $\mu(z)$ because randomisation implies patients not randomised to z give us useful information about the (common) distribution of X
- $\hat{\mu}_2(z)$ is a standardization / G-computation type estimator

Consistency of $\hat{\mu}_2(z)$

- In general, $\hat{\mu}_2(z)$ is only consistent for $\mu(z)$ if the outcome model is correctly specified
- Suppose the outcome model is a canonical GLM. Then the estimation equations are of the form:

$$0 = \sum_{i=1}^{n} \{Y_i - h(X_i, Z_i, \hat{\beta})\} \begin{pmatrix} 1 & X_i & Z_i \end{pmatrix}^T$$

- This implies that the sample mean of the predictions in each treatment group match the sample mean of the outcomes in that treatment group
- It follows for canonical GLMs that $\hat{\mu}_2(z)$ is consistent for $\mu(z)$ even if the model is misspecified
- It is also consistent with negative binomial reg., provided the conditional mean function is correctly specified

Efficiency of $\hat{\mu}_2(z)$

- Using semiparametric theory, one can show that when the outcome model is correctly specified, $\hat{\mu}_2(z)$ is the semi-parametric efficient estimator
- Thus in particular in this case it is more efficient than $\hat{\mu}_1(z)$
- This accords with intuition of using additional information about the common X distribution across all treatment groups due to randomisation

Variance estimation for $\hat{\mu}_2(z)$

- Qu and Luo described a delta method variance estimator $\widehat{\text{Var}}(\hat{\mu}_2(z)|\mathbf{X})$ for $\hat{\mu}_2(z)$ where the target of inference is their previously described alternative parameter
- In most trial settings, the covariates would not be fixed in repeated sampling/trials
- To obtain a variance estimator for $\mu_2(z)$ as an estimator of $\mu(z)$ we can use:

$$\widehat{\text{Var}}(\hat{\mu}_2(z)|\mathbf{X}) + n^{-2} \sum_{i=1}^{n} \{g^{-1}(\hat{\beta}_0 + \hat{\beta}_X^T X_i + \hat{\beta}_Z z) - \hat{\mu}_2(z)\}^2$$

A third estimator

- Existing semiparametric theory for robust covariate adjusted estimation in trials can be used to construct a third estimator:

$$\hat{\mu}_3(z) = \hat{\mu}_1(z) - n^{-1} \sum_{i=1}^{n} \left[\frac{1(Z_i = z) - \hat{\pi}_z}{\hat{\pi}_z} h(X_i, z) \right]$$

  for some function $h(X, z)$
- $\hat{\mu}_3(z)$ is consistent irrespective of choice of $h(X, z)$
- This is because $E(1(Z = z)|X) = \pi_z$, so the added term is always mean zero

Efficiency of $\hat{\mu}_3(z)$

- Efficiency of $\hat{\mu}_3(z)$ is optimised by choosing

$$h(X, z) = E(Y|X, Z = z)$$

- This is of course unknown. We model it, and substitute the prediction:

$$\hat{\mu}_3(z) = \hat{\mu}_1(z) - n^{-1} \sum_{i=1}^{n} \left[\frac{1(Z_i = z) - \hat{\pi}_z}{\hat{\pi}_z} g^{-1}(\hat{\beta}_0 + \hat{\beta}_X^T X_i + \hat{\beta}_Z z) \right]$$

Variance estimation for $\hat{\mu}_3(z)$

- Variance of $\hat{\mu}_3(z)$ can be estimated by

$$\hat{\pi}_z^{-2} n^{-2} \sum_{i=1}^{n} \Big[1(Z_i = z)\{Y_i - \hat{\mu}_3(z)\} \\ - \{1(Z_i = z) - \hat{\pi}_z\}\{h(X_i, z, \hat{\beta}) - \hat{\mu}_2(z)\} \Big]^2$$

Rate estimation

- In some studies we plan to follow patients for a time τ, and Y counts the number of events of a certain type occur for the patient
- A common target of inference is then the rate $E(Y^z)/\tau$
- $\hat{\mu_1}(z)$, $\hat{\mu_2}(z)$ and $\hat{\mu_3}(z)$ readily extend to this setting - see paper

Outline

Motivation

Baseline adjusted mean estimation

Simulations

Conclusions

Simulation setup

- Simulated trials with $n = 400$ patients, two treatments, randomised 1:1
- Single binary baseline covariate X_i
- Follow-up time $T_i = 1$, but for random 25% of patients, $T_i \sim U(0, 1)$
- Event count Y_i then simulated using Poisson under four scenarios:

	True rate	Random effect dist.	Outcome model
1	$\gamma_i \exp(3X_i + Z_i)$	$\gamma_i \sim Ga(2, 0.5)$	Neg. bin.
2	$\gamma_i \exp(3X_i + Z_i)$	$\log(\gamma_i) \sim N(-0.20, 0.41)$	Neg. bin.
3	$\gamma_i \exp(3X_i + Z_i - 1.5X_iZ_i)$	$\gamma_i \sim Ga(2, 0.5)$	Neg. bin.
4	$\gamma_i \exp(3X_i + Z_i - 1.5X_iZ_i)$	$\gamma_i \sim Ga(2, 0.5)$	Poisson

Outcome model always included X_i and Z_i as covariates (but no interaction)

Simulation results

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
$\hat{\mu}_1(z=1)$				
Mean	3.88	3.88	2.80	2.81
95% CI Cov.	94.53	94.28	94.69	94.61
$\hat{\mu}_2(z=1)$				
Bias	0.00	0.00	0.18	0.00
Rel. eff.	1.28	1.28	1.14	1.22
Fixed X CI Cov.	89.61	89.20	81.96	91.15
Random X CI Cov.	94.41	94.20	88.87	95.08
$\hat{\mu}_3(z=1)$				
Bias	0.00	0.00	0.00	0.00
Rel. eff.	1.26	1.25	1.21	1.22
95% CI Cov.	94.47	94.30	94.67	94.56

Outline

Motivation

Baseline adjusted mean estimation

Simulations

Conclusions

Conclusions

- Baseline adjusted mean estimates adjust crude outcome means for observed imbalance in baseline covariates, and have the potential to give more precise estimates
- For certain outcome model types, covariate adjusted estimates are guaranteed to be consistent
- Variance estimation should account for sampling variability in covariates where appropriate
- Contrasts of adjusted marginal mean estimates are identical to adjusted estimates of marginal treatment effects
- See paper for:
  - details for rate estimation
  - impacts of stratified randomisation and missing outcomes
  - illustrative example

More information

- Qu Y, Luo J. Estimation of group means when adjusting for covariates in generalized linear models. Pharmaceutical Statistics, 14(1):56–62, 2015
- Bartlett JW. Covariate adjustment and estimation of mean response in randomised trials. Pharmaceutical Statistics. 2018;1-19. `https://doi.org/10.1002/pst.1880`
- These slides at `www.thestatsgeek.com`
- Simulation code at `www.github.com/jwb133/CovAdjMarginalMean`