Average Marginal Effects in a 2-Arm Parallel Randomized Controlled Trial with Heterogeneity of Effects by Strata

Wade K. Copeland

02 August, 2019

1. Introduction

2. Study Design

3. Average Marginal Treatment Effect Estimation

4. Standard Error of the Average Marginal Treatment Effect

5. Statistical Inference

6. Questions

Section 1

Introduction

Motivation

The general(ized) linear model is very flexible. However, the reliance on effect estimation that requires fixed covariate values can, in some cases, be a severe limitation. This has led statisticians to develop methods related to estimating so-called **average marginal effects**, so named because the method of estimation averages over the marginal effects in a linear model to get a marginal effect that no longer depends on fixed covariate values [Wooldridge, 2000, Greene, 2003]. For the remainder of this presentation we will refer to the **average marginal effect** using the acronym **AME**.

Motivation

The theory behind estimating the AME and calculating the standard error of the estimate is complex. This leads to two general problems in the literature discussing them. The first case abstracts the issue to the point that the average user is unable to decipher what to do in a real application [StataCorp. 2019]. The second case avoids all of the theory and only speaks in general terms, quickly switching to black-box solutions, such as the *margins* command in STATA [Williams, 2012].

Motivation

The utility of the AME in the context of regression modeling makes the lack of accessible resources in the literature it a tragedy for both statisticians and those who consume statistics.

In this presentation, I attempt to solve this problem by deriving the estimate of the AME, and its standard error in the context of a common experimental design; namely, a multisite 2-arm parallel randomized controlled trial (RCT). We follow each section with straight forward programming techniques to apply this method to real data.

Section 2

Study Design

2-Arm Parallel Randomized Control Trial

A 2-arm parallel randomized controlled trial (RCT) is a study design where subjects are assigned randomly to either treatment or control groups (collectively called *condition*). An intervention is applied to the treatment group, and a placebo or standard of care is applied to the control group. At the follow-up, the results between the two groups are compared [Parab & Bhalerao, 2010].

2-Arm Parallel Randomized Control Trial

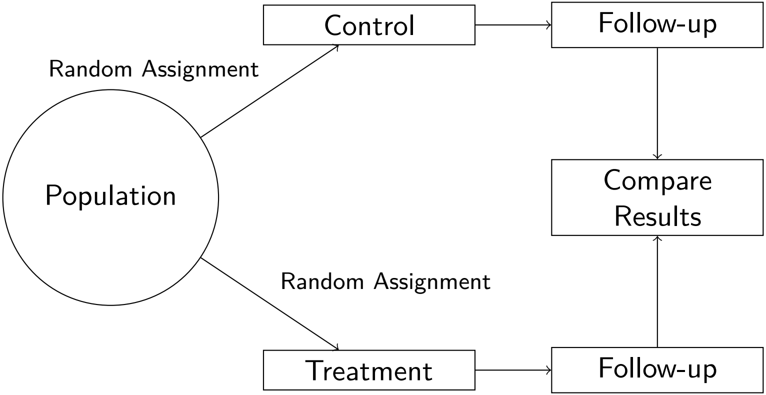

Figure 1: 2-Arm Parallel Randomized Controlled Trial Study Design

Randomization and Causal Effect Estimation

Much of the virtue attributed to the RCT class of experimental designs comes from the process of randomizing participants into either treatment or control groups.

Under ideal circumstances, the randomization allows us to estimate causal effects because unmeasured confounders are balanced across the conditions. This is equivalent to saying that we could have switched subjects from either condition at the start of the study and still ended up with the same results (a property called *exchangeability*[1]) [Hernán & Robins, 2006].

[1] Formally, exchangeability says that the counterfactual outcomes, one outcome observed (the factual outcome) and one that is unobserved (the outcome that would have been observed in a situation that didn't actually happen), are independent of condition.

Multsite RCTs

A multisite RCT is similar only the same experiment is run over multiple sites with each observation within site being randomized (as opposed to randomizing the site such as in a group randomized trial). There are many reasons we would want to do this. Sometimes a single-site RCT is turned into a multisite RCT due to problems at a single location such as low recruitment. A multisite RCT also helps move a trial from only assessing efficacy to assessing the effectiveness by sampling from different populations with different baseline risk factors [Kraemer, 2000].

Marginal Treatment Effect

Statistically, we can account for multiple sites by adjusting for the condition by site effect. The most straight forward way to model the expectation at the follow-up is as a linear function of condition, site, and site by condition.

Marginal Treatment Effect

Consider the following data generating process. Let $y_i = \beta_0 + \beta_1 c_i + \beta_2 s_i + \beta_3 c_i s_i + \epsilon_i$ for $i \in \{1, ...n\}$ such that $y_i \sim N(\beta_0 + \beta_1 c_i + \beta_2 s_i + \beta_3 c_i s_i, \sigma_\epsilon^2)$.

Let $c_i = 0$ for an observation in the control condition and $c_i = 1$ for an observation in the treatment condition. Let $s_i = 0$ for an observation in the first site and $s_i = 1$ for an observation in the second site.

The conditional expectation of y_i for a fixed value of c_i and s_i is as follows:

$$E[y_i|c_i, s_i] = \beta_0 + \beta_1 c_i + \beta_2 s_i + \beta_3 c_i s_i$$

[1] To keep things simple, I assume the conditional distribution of the response is normal. In general, this isn't necessary.

Marginal Treatment Effect

To continue we need to calculate the expected marginal treatment effect for the i^{th} observation.

The effect of the treatment condition is
$E[y_i|c_i = 1, s_i] = \beta_0 + \beta_1 1 + \beta_2 s_i + \beta_3 1 s_i = \beta_0 + \beta_1 + \beta_2 s_i + \beta_3 s_i$

The effect of the control condition is
$E[y_i|c_i = 0, s_i] = \beta_0 + \beta_1 0 + \beta_2 s_i + \beta_3 0 s_i = \beta_0 + \beta_2 s_i$

Therefore the expected marginal treatment effects[2] are the first discrete difference, $E[y_i|c_i = 1, s_i] - E[y_i|c_i = 0, s_i] =$
$\beta_0 + \beta_1 + \beta_2 s_i + \beta_3 s_i - (\beta_0 + \beta_2 s_i) = \beta_1 + \beta_3 s_i$

[2] An important subtlety here is that the marginal treatment effect is calculated for each observation. Let $E[y_{i_{c_i=0}}|s_i]$ be the counterfactual outcome for each observation under the control condition and let $E[y_{i_{c_i=1}}|s_i]$ be the counterfactual outcome for each observation under the treatment condition. Under exchangeability, we can show that
$E[y_{i_{c_i}}|s_i] = E[y_i|c_i, s_i]$ for all $c_i \in \{0, 1\}$[Hernán, 2004].

Section 3

Average Marginal Treatment Effect Estimation

Average Marginal Treatment Effect Estimation

The expected marginal treatment effect for the i^{th} observation, call it $E[y_{d_i}|s_i]$, still depends on value s_i. While the expected treatment effects by site may be of interest, generally we are interested in an effect that is unconditional on site. The basic mechanics of the general(ized) linear model does not allow for this since we can only calculate the marginal expectation for fixed values of site.

Average Marginal Treatment Effect Estimation

A naive approach to the problem would be to test if the interaction of site by condition is significant and then remove it if it isn't. However, this approach falls into the classic trap of saying that failure to reject the null hypothesis is the same as saying there is no effect.

Another potential solution would be to use generalized estimating equations and cluster by site to get the population-averaged treatment effect. However, for those hoping to extend these methods to additional applications[3], they will find this solution doesn't provide the flexibility in estimation they will want.

[3] An example is using the AME to estimate the risk ratio or risk difference for the treatment effect in a logistic regression model. Generalized estimating equations on their own can only estimate the odds ratio for the population-averaged effect. Another example is if we were to use longitudinal models. In this case, we would want to average over the subject-specific effect to account for the correlation and keep site fixed in the expectation, which we can later average over by estimating the AME.

Average Marginal Treatment Effect Estimation

We can solve this problem in a principled way that will provide for the flexibility we will want later. The first step is to apply the Law of Total Expectation to the i^{th} observation of the marginal treatment effect[4] [Bain & Engelhardt, 2000].

$$E_{s_i}[E[y_{d_i}|s_i]] = E[y_{d_i}]$$

[4] Note that we have made the not-so-subtle transition from treating s_i as fixed in the expecatation to s_i as a random variable. Previously we could have treated s_i as random variable and arrived at this same point.

Average Marginal Treatment Effect Estimation

Since s_i partitions the sample space (e.g., $Pr(s_i = 0) + Pr(s_i = 1) = 1$), we can further simplify.

$$\begin{aligned} E[y_{d_i}] = \sum_{j=1}^{2} E[y_{d_i} | s_{ij}] Pr(s_{ij}) = \\ E[y_{d_i} | s_i = 0] Pr(s_i = 0) + E[y_{d_i} | s_i = 1] Pr(s_i = 1) = \\ \beta_1 Pr(s_i = 0) + (\beta_1 + \beta_3) Pr(s_i = 1) = \\ \beta_1 (Pr(s_i = 0) + Pr(s_i = 1)) + \beta_3 Pr(s_i = 1) = \\ \beta_1 + \beta_3 Pr(s_i = 1) \end{aligned}$$

Average Marginal Treatment Effect Estimation

Calculating the unconditional effect for the i^{th} observation is just the conditional marginal treatment effect evaluated at the $Pr(s_i = 1)$ which is the population AME.

Let $y_1, y_2, ...y_n$ be a sequence of n random variables with sample average $\frac{1}{n}\sum_{i=1}^{n} y_{d_i}$. We can see that this is the same as the unconditional marginal treatment effect, and where this estimator gets its name from.

$$\frac{1}{n}\sum_{i=1}^{n} y_{d_i} = \frac{1}{n}\sum_{i=1}^{n}(\beta_1 + \beta_3 s_i) =$$

$$\frac{1}{n}(n\beta_1 + \beta_3\sum_{i=1}^{n} s_i) = \beta_1 + \beta_3\frac{\sum_{i=1}^{n} s_i}{n} =$$

$$\beta_1 + \beta_3 Pr(s_i = 1)$$

Average Marginal Treatment Effect Estimation

This suggests that we can estimate the marginal treatment effect that isn't conditional on site by replacing the population parameters with their consistent estimators. We will call this derived effect $\widehat{AME}$.

$$\widehat{AME} = \hat{\beta}_1 + \hat{\beta}_3 \bar{s} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{d_i}$$

Average Marginal Treatment Effect Estimation – Example

As an example, consider the data generating process with the following values: Let $\beta_0 = 5$, $\beta_1 = 5$, $\beta_2 = 2$, $\beta_3 = 4$ and $\sigma_\epsilon = 1$. Let the probability of an observation being in the treatment condition be $1/2$, and the probability of an observation being in site 1 be $2/3$.

With these values, the population-level effects are easily derived.

1. The population marginal treatment effect in site 1 is $\beta_1 + \beta_3 0 = \beta_1 = 5$.
2. The population marginal treatment effect in site 2 is $\beta_1 + \beta_3 = 5 + 4 = 9$.
3. The population AME is $\beta_1 + \beta_3 Pr(s_i = 1) = 5 + 4(1 - 2/3) = 6.33$.

Average Marginal Treatment Effect Estimation – Example

To see how close $\widehat{AME}$ is to the population AME we can simulate some data.

```
set.seed(123)

beta0 = 5; beta1 = 5; beta2 = 2; beta3 = 4
C = rbinom(500, 1, prob = 1/2)
S = rbinom(500, 1, prob = 1/3)
mu = beta0 + beta1*C + beta2*S + beta3*C*S
Y = rnorm(n = 500, mean = mu, sd = 1)
d <- as.data.frame(cbind(Y, C, S))
d[1:3, ]
```

```
##          Y C S
## 1 4.398107 0 0
## 2 9.006301 1 0
## 3 6.026785 0 0
```

Average Marginal Treatment Effect Estimation – Example

The coefficients show that the estimation is reasonably close to the truth.

```
fit <- lm(Y ~ C + S + C*S, data = d)
coef(fit)
```

```
## (Intercept)           C           S         C:S
##    4.975211    5.040964    2.127431    3.753885
```

The estimated marginal treatment effect at site 1 is $\hat{\beta}_1 + \hat{\beta}_3 0 = 4.98$. The estimated marginal treatment effect at site 2 is $\hat{\beta}_1 + \hat{\beta}_3 = 4.98 + 3.75 = 8.73$. Finally, $\widehat{AME} = \hat{\beta}_1 + \hat{\beta}_3 \bar{S} = 4.98 + 4(0.34) = 6.24$.

Section 4

Standard Error of the Average Marginal Treatment Effect

Standard Error of the Average Marginal Treatment Effect

Estimation is fun and all, but without a method to calculate uncertainty about $\widehat{AME}$, we are up a creek. Fortunately, we are not without many paddles to choose from[Dowd, Greene, & Norton, 2014].

We could try to derive the sampling variance for $\widehat{AME}$ directly, but that might be painful in some cases[5].

A computational approach might be to use Bootstrap Resampling [Efron, 1979], but depending on the end user's computer hardware, this can be slow.

A compromise between accuracy and computation is to use the Delta Method [Doob, 1935]. In short, the Delta Method uses the first two terms of the Taylor power series expansion of the marginal treatment effect evaluated at $\widehat{AME}$ to estimate its variance.

[5] The case in mind is if the expectation is non-linear in the predictors.

Standard Error of the Average Marginal Treatment Effect

The predicted values of y_i are given by the function $f(x_i, \hat{\boldsymbol{\beta}}) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 c_i + \hat{\beta}_2 s_i + \hat{\beta}_3 c_i s_i$, where $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \hat{\beta}_3 & \hat{\beta}_4 \end{bmatrix}^T$.

For fixed values of the c_i and s_i we know that $\hat{\boldsymbol{\beta}}$ corresponds to a consistent estimator of $\boldsymbol{\beta}$. Therefore $\hat{\boldsymbol{\beta}}$ conveges in probability to $\boldsymbol{\beta}$, and by the central limit theorem, converges in distribution (D) to a $N(0, \Sigma)$, where Σ is the variance-covariance matrix of $\boldsymbol{\beta}$.

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(0, \Sigma)$$

Standard Error of the Average Marginal Treatment Effect

Let ∇ be the vector-derivative operator and let g be a scalar-valued function of $\hat{\beta}$. The Delta Method amounts to a generalization of the central limit theorem for any scalar-valued transformation of $\hat{\beta}$. The statement of result is given without proof below.

$$\sqrt{n}(g(\hat{\beta}) - g(\beta)) \xrightarrow{D} N(0, \nabla g(\beta)^T \cdot \Sigma \cdot \nabla g(\beta))$$

To see how this applies to estimating the variance of $\widehat{AME}$ we need to make a couple of observations.

Standard Error of the Average Marginal Treatment Effect

The first observation is that we took the first discrete difference with respect to c_i to transform the $f(x_i, \hat{\boldsymbol{\beta}})$ to a consistent estimator of the predicted marginal treatment effects at the i^{th} site.

$$g(f(x_i, \hat{\boldsymbol{\beta}})) = \hat{\beta}_1 + \hat{\beta}_3 s_i$$

Standard Error of the Average Marginal Treatment Effect

The second observation is that the predicted marginal treatment effects that are unconditional on site corresponds to the predicted marginal treatment effects evaluated at $\bar{s}$.

$$g(f(x_{AME_i}, \hat{\boldsymbol{\beta}})) = \hat{\beta}_1 + \hat{\beta}_3 \bar{s}$$

Therefore $g(f(x_{AME_i}, \hat{\boldsymbol{\beta}}))$ corresponds to a scalar-valued function of our estimators that transforms $f(x_i, \hat{\boldsymbol{\beta}})$ to the marginal treatment effects evaluated at $\bar{s}$.

Standard Error of the Average Marginal Treatment Effect

Applying the Delta Method to g and replacing the population parameters with their estimates, we can derive the approximate variance of $\widehat{AME}$. First note that $\nabla g(f(x_{AME_i}, \hat{\beta}))$ is the vector of partial derivatives with respect to the parameter vector.

$$\left(\frac{\partial g(f(x_{AME_i}, \hat{\beta}))}{\partial f(x_{AME_i}, \hat{\beta})}\right)^T =$$

$$\begin{bmatrix} \frac{\partial g(f(x_{AME_i}, \hat{\beta}))}{\partial \hat{\beta}_0} & \frac{\partial g(f(x_{AME_i}, \hat{\beta}))}{\partial \hat{\beta}_1} & \frac{\partial g(f(x_{AME_i}, \hat{\beta}))}{\partial \hat{\beta}_2} & \frac{\partial g(f(x_{AME_i}, \hat{\beta}))}{\partial \hat{\beta}_3} \end{bmatrix} =$$

$$\begin{bmatrix} 0 & 1 & 0 & \bar{s} \end{bmatrix}$$

Standard Error of the Average Marginal Effect – Example

As an example we will pick up where we left off before. For our data generating process we had $\beta_0 = 5$, $\beta_1 = 5$, $\beta_2 = 2$, $\beta_3 = 4$ and $\sigma_\epsilon = 1$. The probability of an observation being in the treatment condition be $1/2$, and the probability of an observation being in site 1 be $2/3$.

To calculate the standard error of $\widehat{AME}$ we need to know the sample variance-covariance matrix $(\hat{\Sigma})$ and vector-gradient of the function that transforms the predicted values to their marginal treatment effects evaluated at $\widehat{AME}$ $(g(f(x_{AME_i}, \hat{\beta})))$.

Standard Error of the Average Marginal Effect – Example

The sample variance-covariance matrix is easy enough. Note that with our sample size of 500, $\hat{\Sigma}$ is quickly converging to zero.

```
round(vcov(fit), 2)
```

```
##             (Intercept)     C     S   C:S
## (Intercept)        0.01 -0.01 -0.01  0.01
## C                 -0.01  0.01  0.01 -0.01
## S                 -0.01  0.01  0.02 -0.02
## C:S                0.01 -0.01 -0.02  0.04
```

Standard Error of the Average Marginal Effect – Example

The vector-gradient of the function that transforms the predicted values to their marginal treatment effects evaluated at $\widehat{AME}$ is a little harder. Fortunately in R, we don't have to know calculus to symbolically derive the gradient.

Standard Error of the Average Marginal Effect – Example

Let $S = \begin{bmatrix} s_1 & s_2 & \dots s_n \end{bmatrix}^T$ then our expression for the marginal treatment effects is $\hat{\beta}_1 + \hat{\beta}_3 S$.

```
meExpr <- parse(text = "beta_hat_1 + beta_hat_3*S")
meExpr
```

```
## expression(beta_hat_1 + beta_hat_3 * S)
```

Standard Error of the Average Marginal Effect – Example

Using the *deriv* function applied to *meExpr*, we get the gradient matrix for the i^{th} observation (also known as the Jacobian matrix).

```r
meGrad <- deriv(meExpr, c("beta_hat_0", "beta_hat_1",
                          "beta_hat_2", "beta_hat_3"))
meGrad
```

```
## expression({
##     .value <- beta_hat_1 + beta_hat_3 * S
##     .grad <- array(0, c(length(.value), 4L), list(NULL, c("
##         "beta_hat_1", "beta_hat_2", "beta_hat_3")))
##     .grad[, "beta_hat_0"] <- 0
##     .grad[, "beta_hat_1"] <- 1
##     .grad[, "beta_hat_2"] <- 0
##     .grad[, "beta_hat_3"] <- S
##     attr(.value, "gradient") <- .grad
##     .value
## })
```

Standard Error of the Average Marginal Effect – Example

Let $C = \begin{bmatrix} c_1 & c_2 & \dots c_n \end{bmatrix}^T$. The values we need to evaluate the gradient at are set below and then evaluated and stored in the object *J_me*.

```r
beta_hat_1 = coef(fit)["C"]
beta_hat_3 = coef(fit)["C:S"]
S = model.matrix(fit)[, "S"]

J_me <- attr(eval(meGrad), "gradient")
J_me[1:3, ]
```

```
##      beta_hat_0 beta_hat_1 beta_hat_2 beta_hat_3
## [1,]          0          1          0          0
## [2,]          0          1          0          0
## [3,]          0          1          0          0
```

Standard Error of the Average Marginal Effect – Example

The column-wise mean of the Jacobian matrix gives us the vector-gradient for the marginal treatment effects evaluated at $\widehat{AME}$.

```r
J_ame <- t(matrix(apply(J_me, 2, mean)))
J_ame
```

```
##      [,1] [,2] [,3]  [,4]
## [1,]    0    1    0 0.338
```

Standard Error of the Average Marginal Effect – Example

Putting it all together the standard error of $\widehat{AME}$ is as follows:

```r
ame_se <- sqrt(J_ame %*% vcov(fit) %*% t(J_ame))
ame_se
```

```
##             [,1]
## [1,] 0.09069116
```

Section 5

Statistical Inference

Statistical Inference

The final piece of the puzzle is to apply inferential statistics to test the null hypothesis of no average marginal treatment effect and calculate confidence intervals.

We can test the null hypothesis, $H_0 : \widehat{AME} = 0$, by using a simple Z-test.

$$Z_{obs} = \frac{\widehat{AME} - \widehat{AME}_{H_0}}{SE(\widehat{AME})} = \frac{\widehat{AME}}{SE(\widehat{AME})}$$

Under a true null $Z \sim N(0, 1)$. Therefore under the assumption that the null hypothesis of no average marginal treatment effect is true, the P-value is two times the upper tail probability of observing Z_{obs} or greater. The 95% confidence interval for $\widehat{AME}$ is then:

$$\widehat{AME} \pm Z_{.975} SE(\widehat{AME}) = \widehat{AME} \pm 1.96 SE(\widehat{AME})$$

Statistical Inference – Example

To show statistical inference in application we continue where we left off.

The p-value is calculated below and shows that we can reject the null hypothesis of no average marginal treatment effect with gusto!

```r
z <- (coef(fit)[1] + coef(fit)[4]*mean(d$S))/ame_se
p <- 2*pnorm(abs(z), lower.tail = FALSE)
p
```

```
##      [,1]
## [1,]    0
```

Statistical Inference – Example

The lower and upper 95% confidence limits are calculated below. The results show that there is good evidence that the population AME lies between the bounds 6.07 and 6.42[6].

```r
ci_lb <- (coef(fit)[1] + coef(fit)[4]*mean(d$S)) -
  1.96*ame_se
ci_ub <- (coef(fit)[1] + coef(fit)[4]*mean(d$S)) +
  1.96*ame_se

paste("[", round(ci_lb, 2), ", ", round(ci_ub, 2), "]",
      sep = "")
```

```
## [1] "[6.07, 6.42]"
```

[6] Specifically, if we repeat the experiment many times, calculating the confidence interval each time, we would expect them to contain the population AME 95% of the time.

Section 6

Questions

References

Bain, Lee J and Engelhardt, Max, 2000
Introduction to probability and mathematical statistics
Brooks/Cole.

Doob, J. L., 1935
The Limiting Distributions of Certain Statistics
Annals of Mathematical Statistics, 6, 160—169.

Dowd, Bryan E and Greene, William H and Norton, Edward C, 2014
Computation of standard errors
Health services research, 49(2), 731-750.

Efron, Bradley, 1979
Bootstrap methods: another look at the jackknife
Annals of Statistics, 7, 1–26.

References

Greene, William H, 2003
Econometric analysis
Pearson Education India.

Hernán, Miguel Angel, 2004
A definition of causal effect for epidemiological research
Journal of Epidemiology & Community Health 58(4), 265-271.

Hernán, Miguel A and Robins, James M, 2006
Estimating causal effects from epidemiological data
Journal of Epidemiology & Community Health 60(7), 578–586.

Kraemer, Helena Chmura, 2000
Pitfalls of multisite randomized clinical trials of efficacy and effectiveness
Schizophrenia Bulletin 26(3), 533–541.

References

Parab, Shraddha and Bhalerao, Supriya, 2010
Study designs
International journal of Ayurveda research 1(2), 128.

R Core Team, 2019
R: A language and environment for statistical computing
R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/.

StataCorp, 2019
Stata 16 Base Reference Manual
College Station, TX: Stata Press.

Williams, Richard, 2012
Using the margins command to estimate and interpret adjusted predictions and marginal effects
The Stata Journal 12(2), 308–331.

Wooldridge, Jeffrey M, 2000
Econometric analysis of cross section and panel data
MIT press.