VR and Regressin Adjustment

Alex Deng

1/22/2019

Contents

1 VR and Regressin Adjustment **1**

1.1 Variance Reduction 1

1 VR and Regressin Adjustment

1.1 Variance Reduction

So far we have been using the simple difference of sample averages Δ to estimate the ATE δ. Randomization guarantees it is unbiased, that is $E(\Delta) = \delta$. The statistical power of the two sample test based on Δ relies on its variance $\mathrm{Var}(\Delta)$. Reducing its variance increases statistical power and the sensitivity of the metric.

Usually, there is a trade-off between variance and bias, so called bias-variance trade-off. However, as we will see, there is a way to reduce the variance without introducing any bias. Statisticians calls this efficiency augmentation (Tsiatis 2006). Efficiency augmentation here means more accurate estimation using the same amount of information or sample size. In other words, we seek to come up with a new estimator Δ^* to replace Δ that is still unbiased and with smaller variance. We show a simple and powerful idea based on variance reduction using control variates from Monte Carlo simulation. The general theory of regression adjustment and semiparametric efficiency augmentation is closely related to the idea of doubly robust estimation (Kang and Schafer 2007).

1.1.1 Control Variates and CUPED

In the context of Monte Carlo simulation, we face a similar problem of estimate the mean $\mathrm{E}(Y)$ of a random variable Y for which we can simulate (draw sample) from. The naive Monte Carlo estimator is the sample mean $\overline{Y}$, similar to the naive ATE estimator Δ. Control variates provide an alternative Monte Carlo estimator with smaller variance. To do that, we need another random variable X with *known* mean $\mu_x = \mathrm{E}(X)$. For any fixed value of θ, the following is also an unbiased estimator for $\mathrm{E}(Y)$:

$$\widehat{Y}_{cv} := \overline{Y} - \theta\overline{X} + \theta\mu_x \,.$$

The unbiasedness of $\widehat{Y}_{cv}$ is due to the fact that last two terms on the right hand side cancels with each other since $\mathrm{E}(\overline{X}) = \mu_x$. Also note that this estimator requires μ_x to be known to even be defined.

The variance of this newly created estimator $\widehat{Y}_{cv}$ is

$$\begin{aligned}\mathrm{Var}(\widehat{Y}_{cv}) &= \mathrm{Var}(\overline{Y} - \theta\overline{X}) = \mathrm{Var}(Y - \theta X)/n \\ &= \frac{1}{n}(\mathrm{Var}Y) + \theta^2\mathrm{Var}(X) - 2\theta\mathrm{Cov}(Y, X)) \,.\end{aligned}$$

Note that $\mathrm{Var}(\widehat{Y}_{cv})$ is minimized when we choose

$$\theta = \mathrm{Cov}(Y, X)/\mathrm{Var}(X) \tag{1}$$

and with this optimal choice of θ, we have

$$\mathrm{Var}(\widehat{Y}_{cv}) = \mathrm{Var}(\overline{Y})(1-\rho^2)\,,$$

where $\rho = \mathrm{Cor}(Y, X)$ is the *correlation* between Y and X. That is

$$\frac{\mathrm{Var}(\widehat{Y}_{cv})}{\mathrm{Var}(\overline{Y})} = 1-\rho^2\,.$$

That is, the variance is reduced by a factor of ρ^2! The larger ρ, the better the variance reduction.

The single control variate case can be easily generalized to include multiple variables. It is interesting to point out the connection with linear regression. The optimal θ turns out to be the ordinary least square (OLS) solution of regressing (centered) Y on (centered) X, which in multiple variable case has variance

$$\mathrm{Var}(\widehat{Y}_{cv}) = \mathrm{Var}(\overline{Y})(1-R^2)\,,$$

with R^2 being the proportion of variance explained coefficient from the linear regression.

It is also possible to use nonlinear adjustment. Instead of allowing only linear adjustment, we can minimize variance in a more general functional space. Let

$$\widehat{Y}_{cv} = \overline{Y} - \overline{f(X)} + \mathrm{E}(f(X)), \tag{2}$$

and then try to minimize the variance of (2). It can be shown that the regression function $\mathrm{E}(Y|X)$ gives the optimal $f(X)$.

Exercise: Prove $f(X) = \mathrm{E}(Y|X)$ is the optimal control variates for Y using X.

So far we have been looking at one sample mean and assume the mean of control variate μ_x to be known. Utilizing control variates to reduce variance is a very common technique in Monte Carlo simulation (Asmussen and Glynn 2008). The difficulty of applying it usually boils down to finding a control variate X that is highly correlated with Y and at the same time has known $\mathrm{E}(X)$.

Deng et al. (2013) made the observation that in a randomized experiment, we don't need to know μ_x to use X as control variate because we care about the ATE, not the two means $\mathrm{Y}^{(\mathrm{t})}$ and $\mathrm{Y}^{(\mathrm{c})}$ for treatment and control groups. If we replace $\overline{Y^{(t)}}$ by $\overline{Y_{cv}^{(t)}}$ and $\overline{Y^{(c)}}$ by $\overline{Y_{cv}^{(c)}}$, and then define

$$\begin{aligned}\Delta^* &:= \overline{Y_{cv}^{(t)}} - \overline{Y_{cv}^{(c)}} \\ &= \Delta(Y) - \theta\Delta(X) + \theta(\mathrm{E}(X^t) - \mathrm{E}(X^c))\,.\end{aligned}$$

Here $\Delta(Y) = \overline{Y^t} - \overline{Y^c}$ and $\Delta(X) = \overline{X^t} - \overline{X^c}$ are simple difference of sample means. From Equation (3) it is apparent that if $\mathrm{E}(X^t) = \mathrm{E}(X^c))$, the last term disappear and

$$\Delta^* = \Delta(Y) - \theta\Delta(X)\,. \tag{3}$$

In a few elementary steps, we have achieved wonder. This new Δ^* in Equation (3) does not involve any unknown mean $\mathrm{E}(X^t)$ or $\mathrm{E}(X^c)$. Moreover, its variance can be greatly reduced comparing to the original ATE estimator Δ thanks to control variate X. The only requirement is the control variates X we picked need to satisfy the condition

$$\mathrm{E}(X^t) = \mathrm{E}(X^c)\,.$$

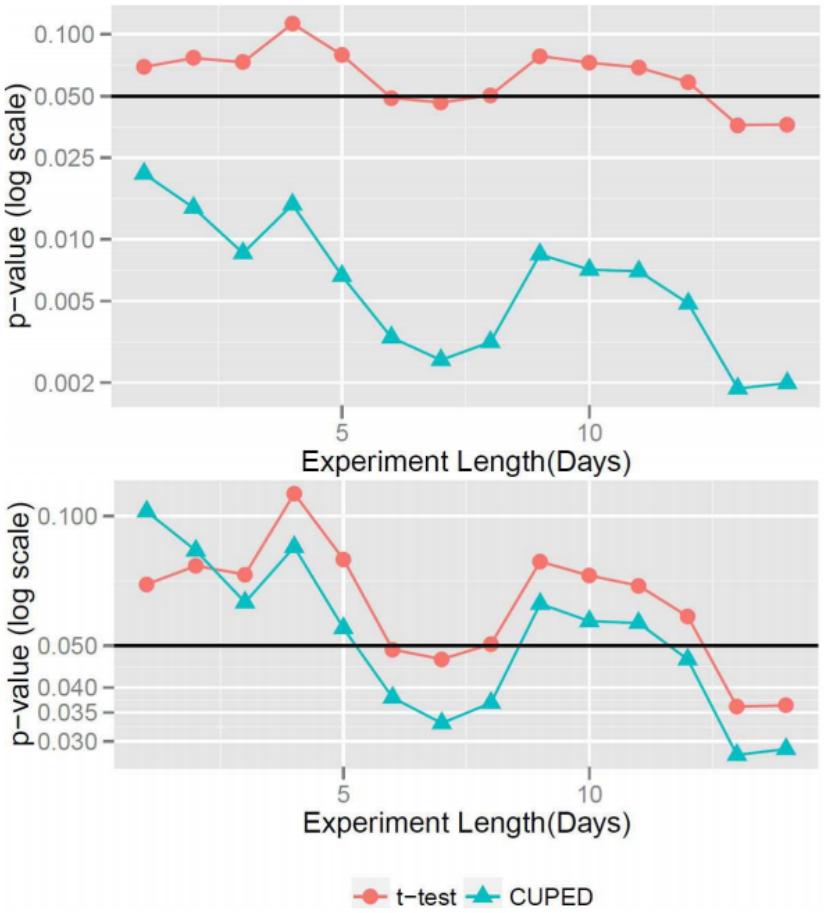

Figure 1: Variance Reduction in Action for a real experiment. Top: p-value. Bottom: p-value when using only half the users for CUPED.

that is, the ATE on X must be zero. These kind of X are abundant in practice, because a treatment cannot possibly impact anything observed **before** an experiment unit get exposed to the treatment. Deng et al. (2013) calls these *pre-experiment* variables and named the simple procedure described so far *CUPED* (Controlled experiment Using Pre-Experiment Data), and demonstrated the effectiveness of using the same metric data for the pre-experiment period as control variates performs well in practice. Figure 1 shows the variance of a metric reduced by more than 50%. With only half of the original sample size, Δ^* of CUPED produces more statistical power than estimating ATE by Δ.

A few important remarks:

1. Optimal θ is $\mathrm{Cov}(Y, X)/\mathrm{Var}(X)$ for control variates. In CUPED, we have treatment and control groups. Should we use treatment or control group to define the optimal θ? The answer is it usually does not matter much. Note that the CUPED estimator Δ^* is unbiased for *any* fixed value of θ, and different θ merely affect variance reduction rate. As long as the treatment effect is not too big, the optimal θ for the two groups are very close and it does not make a lot of difference which one to choose. If needed, one can optimize θ to minimize the variance of Δ^* directly. The minimizer of $\mathrm{Var}(\Delta^*)$ is

$$\frac{\mathrm{Cov}(\overline{Y^t}, \overline{X^t}) + \mathrm{Cov}(\overline{Y^c}, \overline{X^c})}{\mathrm{Var}(\overline{X^t}) + \mathrm{Var}(\overline{X^c})}. \tag{4}$$

   Another choice is to use pooled data of treatment and control to compute θ. In Section 1.1.2 we will

use results from more general semiparametric theory to paint a much clearer picture.

2. Equation (3) can be trivially extended to nonlinear adjustment:

$$\Delta^* := \overline{Y_{cv}^{(t)}} - \overline{Y_{cv}^{(c)}} = \Delta(Y) - \Delta(f(X)) \,. \tag{5}$$

   Similar to control variates, the closer $f(X)$ to $\mathrm{E}(Y|X)$, the better the variance reduction.
3. Pre-experiment data does not literally means data collected before the experiment begins. It can be anything before the *triggering* of the treatment intervention. For example, the day-of-week a user is first observed in the experiment is independent of the experiment itself, as well as the age, gender, browser and the device a user uses.
4. Control variates X can be categorical (discrete) or continuous. When X is categorical, we can use one-hot encoder to create dummy binary variables, and CUPED can be seen as post-stratification adjustment, which is shown to be asymptotically as efficient as doing actual blocking (stratified sampling) (Miratrix, Sekhon, and Yu 2013).

5. For some choice of X, it might be not be well-defined for a subset of experiment units. For example, new users just appeared during the experiment do not exist before the experiment period and the pre-experiment period metric value for them are simply not defined. In such cases, Deng et al. (2013) proposed to impute with 0 and at the same time also include an binary indicator variate to indicate whether a unit has valid pre-experiment and use both in a multiple regression version of CUPED.
6. There is a deep connection between control variates and linear regression. However, control variates method does not actually assume any linear relationship between X and Y. The linear regression and the optimal θ being solution of OLS is simply a *working model*. The extension to nonlinear $f(X)$ makes it clear that the working model can be any model and the quality of the model only affects the variance reduction rate. In the next section we will use a more general semiparametric model to emphasize the distinction and show CUPED is quite general and include two regression models as special cases.

1.1.2 General Semiparametric Regression Adjustment

There is a clear resemblance between CUPED and linear regression. Let A_i be the binary treatment assignment indicator of the ith experiment unit. Consider the following two common linear regression models:

$$Y_i = \alpha + \delta A_i + \beta X_i + \epsilon_i \,, \tag{6}$$

and

$$Y_i = \alpha + \delta A_i + \beta X_i + (\gamma X_i) A_i \epsilon_i \,. \tag{7}$$

Under the standard linear regression model assumptions, the regressors X and A are considered to be fixed, the residual ϵ_i are assumed to be i.i.d. with a normal distribution. The only random component in the underlying data-generating-process are from the residuals alone. $\alpha + \beta X$ is the prediction for Y in the control group based on X. δ is the average treatment effect and γX_i in the second model is the linear treatment effect adjustment that allows the conditional treatment effect $\mathrm{E}(\tau|X)$ to be a linear function of X. Fitting the linear model to get estimators of those coefficients and their sampling distributions are also known as *Analysis of (Co)Variances* (ANOVA/ANCOVA). We call the first model *ANCOVA1* and the second *ANCOVA2*.

Both models are widely used in two group comparison for both experiment data and also observational data. Many have pointed out the *efficiency gain* from the regression model to increase accuracy of estimating the average treatment effect δ. Nevertheless, it is obvious that the linear model is too restrictive. The data-generating-process for real world problems will involve random X and A; the true regression $\mathrm{E}(Y|X, A)$ will unlikely be linear, so ϵ may not even satisfy $\mathrm{E}(\epsilon|X) = 0$, let alone i.i.d. normally distributed.

Freedman (Freedman 2008) criticized the practices of using parametric linear regression theory on experimentation data, stating: "randomization does not justify the models, bias is likely; nor are the usual variance calculations to be trusted." Using Neyman's complete randomization with potential outcomes, Freedman avoid postulating a parametric model for $(Y(T), Y(C), X)$ by treating them as fixed and the only random

component is the treatment assignment A. Asymptotic and finite sample theories for the *ANCOVA1* model was given in Freedman (2008); and a following work Lin (2013) studied the *ANCOVA2* model.

Here we follow the independent randomization model and assume (Y, X, A) are independently sampled from a super population. The model we use is general. The joint distribution of (Y, X, A) are allowed to be anything except the restriction that A is result of independent randomization. That is, the joint density has a natural decomposition as

$$p(y, x, a) = p_y(y|x, a)p_a(a|x)p(x) \tag{8}$$

and $p_a(a|x)$ is known to be the Bernoulli density $p^a(1-p)^{(1-a)}$ with fixed treatment probability p. This kind of model with a combination of both parametric and nonparametric components are called *semiparametric model*.

The target of the inference is to estimate the ATE

$$\delta = \mathrm{E}(Y|A = 1) - \mathrm{E}(Y|A = 0)\,.$$

Unlike in a parametric model where the target of inference is either one of the model parameters or a function of them, for semiparametric model, the inference can be any functional of the distribution.

For large sample, asymptotic theories exist for semiparametric model just like parametric model. It can be shown that all reasonable consistent and asymptotically normal estimators for δ are either exactly or asymptotically equivalent to this form:

$$\overline{Y^{(t)}} - \overline{Y^{(c)}} + \frac{1}{n}\sum \big((A_i - p)h(X_i)\big) \tag{9}$$

for a function h of X.

A rigorous explanation is beyond our scope and can be found in Tsiatis (2006), Van der Vaart (2000) or Robins and Rotnitzky (1995). Here we just state some general results focusing on high level intuitions. Asymptotic theory for semiparametric models focus on only regular and asymptotically linear estimators (RAL). Regularity condition is to avoid pathological estimators whose behavior can vary dramatically in the neighborhood of the ground truth value, as exemplified by Hodges' estimator (Van der Vaart 2000). Consistent RAL estimators represents all reasonable estimators of interest with nice properties such as asymptotical normality, including MLE for parametric models, M-estimator and Z-estimator.

Semiparametric theory guarantees that all consistent RAL estimators are asymptotically equivalent to an estimator of the form

$$\overline{\psi(Y, X, A)} + \overline{h(Y, X, A)}\,,$$

where $\overline{\psi(Y, X, A)}$ is *any* consistent RAL estimator and $h(Y, X, A)$ is from a linear subspace of the Hilbert space of mean-zero finite variance random functions. This linear subspace, denoted by $\mathcal{T}^\perp$, is the orthogonal component of the *tangent space* $\mathcal{T}$. The tangent space for a parametric model can be defined as the linear subspace spanned by score functions. For a semiparametric model, the tangent space contains the tangent space of any parametric submodel – that is, a parametric model whose distribution is also included in the semiparametric model.

For our purpose, we already have a consistent RAL estimator. The naive Δ estimator $\overline{Y^{(t)}} - \overline{Y^{(c)}}$ is asymptotically equivalent to

$$\overline{\frac{AY}{p}} - \overline{\frac{(1-A)Y}{1-p}}\,.$$

Turns out that the linear space $\mathcal{T}^\perp$ has a very simple form. It contains all mean-zero finite variance functions $f(A, X)$ such that

$$\mathrm{E}(f(A, X)|X) = 0\,.$$

Since $f(A, X)$ is $f(1, X)$ with probability p and $f(0, X)$ with probability $1-p$, the above condition together with the independence of A and X entails

$$f(A, X) = (A - p)f(1, A)\,.$$

Let $h(X) = f(1, A)$, we have shown Equation (9) characterizes all consistent RAL estimators for ATE δ.

Exercise: Show $\overline{Y^{(t)}} - \overline{Y^{(c)}}$ is asymptotically equivalent to $\overline{\frac{AY}{p} - \frac{(1-A)Y}{1-p}}$. Show $f(A, X) = (A-p)f(1, A)$ if $\mathrm{E}(f(A, X)|X) = 0$ and A is independent Bernoulli(p).

Because $\mathrm{E}\left((A_i - p)h(X_i)\right) = 0$, Equation (9) can be seen as a sum of any consistent RAL estimator and another estimator of 0. This is similar to CUPED where we augment naive ATE estimator Δ by $\theta\Delta(X)$. Like in CUPED we optimize θ to minimize variance, here we can minimize the variance of (9) to find the optimal functional form of $h(X)$.

Using the asymptotic equivalent form $\overline{\frac{AY}{p} - \frac{(1-A)Y}{1-p}}$ of $\overline{Y^{(t)}} - \overline{Y^{(c)}}$, minimize the variance of Equation (9) is to minimize variance of

$$\frac{AY}{p} - \frac{(1-A)Y}{1-p} + (A-p)h(X)\,,$$

which is attained if and only if

$$\mathrm{E}\left[\left(\frac{AY}{p} - \frac{(1-A)Y}{1-p} + (A-p)h(X)\right) \times (A-p)g(X)\right] = 0 \quad \text{for any } g(X).$$

Let $h_1(X) = \mathrm{E}(Y|X, A=1)$ and $h_0(X) = \mathrm{E}(Y|X, A=0)$,

$$\begin{aligned}\mathrm{E}\left[\frac{AY}{p}(A-p)g(X)\right] &= \mathrm{E}\left[\frac{AY}{p}(A-p)g(X)|X\right] \\ &= \mathrm{E}[(1-p)h_1(X)g(X)]\,.\end{aligned}$$

Similarly,

$$\mathrm{E}\left[\frac{(1-A)Y}{1-p}(A-p)g(X)\right] = \mathrm{E}[ph_0(X)g(X)]\,, \tag{10}$$

$$\mathrm{E}[(A-p)^2 h(X)g(X)] = p(1-p)\mathrm{E}[h(X)g(X)]\,. \tag{11}$$

We need

$$\mathrm{E}\left[(1-p)h_1(X)g(X) + ph_0(X)g(X) + p(1-p)h(X)g(X)\right] = 0 \quad \text{for any } g(X)$$

and it can only happen if

$$h(X) = -\frac{h_1(X)}{p} - \frac{h_0(X)}{1-p}\,.$$

Equation (9) with this optimal augmentation becomes

$$\overline{Y^{(t)}} - \overline{Y^{(c)}} - \frac{1}{n}\sum\left((A_i - p)\left(\frac{h_1(X)}{p} + \frac{h_0(X)}{1-p}\right)\right) \tag{12}$$

which is also asymptotically equivalent to

$$\overline{Y^{(t)}} - \overline{Y^{(c)}} - \sum\left((A_i - \overline{A})\left(\frac{h_1(X)}{n_t} + \frac{h_0(X)}{n_c}\right)\right) \tag{13}$$

Estimator (13) solves the problem of the most efficient consistent and RAL estimator for the ATE δ under the semiparametric model where we make not a single model assumption other than the independent randomization. To use that, we need to know the true regression $h_1(X) = \mathrm{E}(Y|X, A=1)$ and $h_0(X) = \mathrm{E}(Y|X, A=0)$ which requires separated works. However, any choice of h_1 and h_0 in (13) is still of the form (9) and is a consistent RAL estimator. In particular, when $h_0(X) = h_1(X) = 0$, estimator (13) reduced to Δ. When $h_0(X) = h_1(X) = f(X)$, estimator (13) reduces to

$$\overline{Y^{(t)}} - \overline{Y^{(c)}} - (\overline{f(X^{(t)})} - \overline{f(X^{(c)})})$$

which is the same as general CUPED in Equation (5).

Exercise: Show Equation (13) reduce to Equation (5) when $h_0(X) = h_0(X) = f(X)$.

Tsiatis et al. (2008) showed both *ANCOVA1* and *ANCOVA2* are special case of estimators of the form (13) when $h_0(X) = h_1(X) = f(X)$. From this perspective, they are also special cases of CUPED estimator (5). The differences between *ANCOVA1* and *ANCOVA2* are the choice of how to fit linear regression $f(X)$ using treatment and control data. Recall that the linear coefficient estimator is $\text{Cov}(X,Y)/\text{Var}(X)$. the denominator $\text{Var}(X)$ is the same for treatment and control. $\text{Cov}(X,Y)$ are different due to the treatment effect. *ANCOVA1* pools the data together and fit a linear regression. This is to use

$$\text{Cov}(X,Y) = p\text{Cov}_T(X,Y) + (1-p)\text{Cov}_C(X,Y)\,,$$

where $\text{Cov}_T(X,Y)$ is the covariance in the treatment, $\text{Cov}_C(X,Y)$ for control and p is the proportion of treatment sample size. On the contrary, *ANCOVA2* uses

$$\text{Cov}(X,Y) = (1-p)\text{Cov}_T(X,Y) + p\text{Cov}_C(X,Y)\,.$$

For CUPED, in the end of Section 1.1.1 we showed if we optimize θ to minimize the variance of CUPED estimator Δ^* directly, the optimal θ is

$$\frac{\text{Cov}(\overline{Y^t},\overline{X^t}) + \text{Cov}(\overline{Y^c},\overline{X^c})}{\text{Var}(\overline{X^t}) + \text{Var}(\overline{X^c})}\,.$$

This θ asymptotically converge to $(1-p)\text{Cov}_T(X,Y) + p\text{Cov}_C(X,Y)$. Hence CUPED with this arrangement of θ is asymptotically equivalent to *ANCOVA2*.

Because $\text{Cov}(\overline{Y^t},\overline{X^t}) = \text{Cov}_T(X,Y)/n_t$ and $\text{Cov}(\overline{Y^c},\overline{X^c}) = \text{Cov}_C(X,Y)/n_c$, we can see covariances are weighted inversely proportional to the sample sizes. This explains why it is better to weight $\text{Cov}_T(X,Y)$ by $1-p$ and $\text{Cov}_C(X,Y)$ by p, instead of using a more straightforward choice of p for $\text{Cov}_T(X,Y)$ and $1-p$ for $\text{Cov}_C(X,Y)$ as in *ANCOVA1* . From here we also see *ANCOVA2* is theoretically better than *ANCOVA1*. Although in practice this difference is small unless p is away from 0.5 and $\text{Cov}_T(X,Y)$ is very different from $\text{Cov}_C(X,Y)$.

1.1.3 Doubly Robust Estimator

Before we close the discussion of regression adjustment, take a look at the doubly robust estimator. DR estiator when propensity score is known is the same as (13)! The goal of doubly robust estimator was to combine regression prediction to impute missing counterfactuals with propensity reweighting so as long as one of the two models is unbiased the DR estimator remains unbiased. For randomized experiments the propensity model is known and the DR estimation is unbiased for arbitrary regression model. This is the spirit of regression adjustment in (13). The closer the regression model is to the true regression $h_0 = \text{E}(Y|X, A = 0)$ and $h_1 = \text{E}(Y|X, A = 1)$, the better (smaller variance).

Asmussen, Soren, and Peter Glynn. 2008. *Stochastic Simulation*. Springer-Verlag.

Deng, Alex, Ya Xu, Ron Kohavi, and Toby Walker. 2013. "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data." In *Proc. 6th Acm Int. Conf. Web Search Data Min.*, edited by acm, 123–32. ACM.

Freedman, David A. 2008. "On regression adjustments to experimental data." *Adv. Appl. Math.* 40 (2). Elsevier: 180–93.

Kang, Joseph DY, and Joseph L Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science*. JSTOR, 523–39.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *Ann. Appl. Stat.* 7 (1). Institute of Mathematical Statistics: 295–318.

Miratrix, Luke W, Jasjeet S Sekhon, and Bin Yu. 2013. "Adjusting treatment effect estimates by post-stratification in randomized experiments." *J. R. Stat. Soc. Ser. B (Statistical Methodol.* 75 (2). Wiley Online Library: 369–96.

Robins, James M, and Andrea Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90 (429). Taylor & Francis Group: 122–29.

Tsiatis, Anastasios A. 2006. *Semiparametric Theory and Missing Data*. Springer-Verlag.

Tsiatis, Anastasios A., Marie Davidian, Min Zhang, and Xiaomin Lu. 2008. "Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach." *Stat. Med.* 27.

Van der Vaart, Aad W. 2000. *Asymptotic statistics*. Vol. 3. Cambridge university press.