Latent Stratification for Incrementality Experiments

Ron Berman and Elea McDonnell Feit

Online Appendix

OA.1 Other applications of principal stratification

Latent stratification is a novel adaptation of principal stratification (Frangakis and Rubin 2002) that is well-suited to advertising experiments where many outcomes are zero. In principal stratification for treatment non-compliance (Imbens and Rubin 1997), the strata represent potential outcomes for treatment compliance. That is, the strata are defined by whether the treatment each unit actually received (Z') match the randomly assigned treatment (Z). The objective is to understand the effect of the applied treatment, which Imbens and Rubin dub the complier average causal effect (CACE). This is distinct from latent stratification which does not address treatment non-compliance. See Table OA.1.1.

Latent stratification is more similar to truncation-by-death (Zhang et al. 2009), where the outcome is undefined for people who have died, e.g. heart rate for an individual who had died before the study endpoint. The strata are defined by whether or not the outcome Y is observed under treatment and control. The objective is to estimate the treatment effect for the A stratum. Treatment effects for other strata and the overall ATE are undefined (Zhang et al. 2009).

Latent stratification is different than these other applications because it stratifies based on whether the customer makes a purchase. Our first insight is that not-purchasing has a defined value of zero and we can compute a stratified overall ATE as a weighted average of the strata ATEs. Other applications of principal stratification do not estimate an overall ATE. Our second insight is that the overall ATE has lower sampling variance than the standard difference-in-means. The overall ATE is computed as a weighted average of the observed strata. Computing treatment effects for observed strata and then averaging them together is called post-stratification and can reduce sampling variance (Miratrix et al. 2013). The key difference between our approach and standard post-stratification is that the strata are latent: thus the name "Latent Stratification".

Table OA.1.1: Key differences between applications of principal stratification

Application	Strata defined by potential outcomes for:	Goal is to infer:
Treatment Non-compliance	$Z' = Z$	ATE for stratum B ("CACE")
Truncation by Death	$Y = \text{NA}$	ATE for stratum A
Latent Stratification	$Y = 0$	Overall ATE

OA.2 Gradient and Hessian R code

The analytical expressions for the observed gradient and Hessian of the latent stratification model can be easily derived using software such as Mathematica by differentiating the log-likelihood directly. Because the resulting expressions are quite long and complex, we include R code for computing these values in the functions `gr_ll_ls()` and `hes_ll_ls()` for the gradient and the Hessian, respectively.

```
# Compute the gradient of the log-likelihood for the latent stratification model.
#
# par is the vector c(piA, piB, muA1, muA0, muB1, sigma).
# dt is a data frame containing cols y (outcome),
# x (non-zero outcome indicator), z (treatment indicator)
# and xz (TRUE if observation is both treated and has non-zero y).
#
# Returns the gradient vector with a size of length(par)=6 items.
#
gr_ll_ls <- function(par, dt) {
  piA <- par[1]
  piB <- par[2]
  piC <- 1 - piA - piB
  muA1 <- par[3]
  muA0 <- par[4]
  muB1 <- par[5]
  sigma <- par[6]

  if (piA < 0 | piB < 0 | piC < 0 | sigma < 0)
    stop("Error in gr_ll_ls(): piA, piB, piC or sigma < 0")

  y <- dt$y
  z <- dt$z
  x <- dt$x
  xz <- dt$xz

  y_xz <- y[xz]

  # compute Normal density for positive, treated (xz) observations
  fA1 <- dnorm(y_xz, muA1, sigma) # only compute for "treated, purchase" group
```

```
  fB1 <- dnorm(y_xz, muB1, sigma)
  if (sum(is.infinite(c(fA1, fB1))))
    warning("Numeric overrun in Normal density calculation in gr_ll_ls()")

  # partial derivatives
  dpiA <- sum( fA1/(piA*fA1 + piB*fB1) ) - sum( (1-x)*z )/(1-piA-piB) +
    sum(x*(1-z))/piA - sum((1-x)*(1-z))/(1-piA)
  dpiB <- sum( fB1/(piA*fA1 + piB*fB1) ) - sum( (1-x)*z )/(1-piA-piB)
  dmuA1 <- sum( (piA*fA1/(piA*fA1 + piB*fB1))*((y_xz-muA1)/sigma^2) )
  dmuA0 <- sum( x*(1-z)*(y-muA0)/sigma^2 )
  dmuB1 <- sum( (piB*fB1/(piA*fA1 + piB*fB1))*((y_xz-muB1)/sigma^2) )
  dsigma <- sum( (piA*fA1*((y_xz-muA1)^2-sigma^2)/sigma^3 +
    piB*fB1*((y_xz-muB1)^2-sigma^2)/sigma^3 )/(piA*fA1 + piB*fB1) ) +
    sum( x*(1-z)*((y-muA0)^2-sigma^2)/sigma^3 )

  out <- c(piA=dpiA, piB=dpiB, muA1=dmuA1, muA0=dmuA0, muB1=dmuB1, sigma=dsigma)
}

# Computes the Hessian (matrix of second derivatives)
# of the log-likelihood for the latent stratification model.
# dt is a data frame containing cols y (outcome),
# x (non-zero outcome indicator) and z (treatment indicator).
#
# Returns the a 6x6 matrix of second derivatives.
#
hes_ll_ls <- function(par, dt) {
  piA <- par[1]
  piB <- par[2]
  piC <- 1 - piA - piB
  muA1 <- par[3]
  muA0 <- par[4]
  muB1 <- par[5]
  s <- par[6]

  y <- dt$y
  z <- dt$z
  x <- dt$x

  if (piA < 0 | piB < 0 | piC < 0 | s < 0)
    stop("Error in gr_ll_ls(): piA, piB, piC or sigma < 0")

  hes <- matrix(NA, nrow=6, ncol=6)
  Q3 <- exp((muB1-y)^2/(2*s^2))
  Q4 <- exp((muA1-y)^2/(2*s^2))
  Q1 <- (piA*Q3 + piB*Q4)^2 # (piA*exp((muB1-y)^2/(2*s^2))+piB*exp((muA1-y)^2/(2*s^2)))^2
  Q2 <- Q3*Q4 # exp( (muA1^2+muB1^2-2*y*(muA1+muB1)+2*y^2) / (2*s^2) )
  hes[1,1] <- -sum((1-x)*(1-z) / (piA-1)^2)-
    sum((1-x)*z/(piA + piB - 1)^2)-
    sum(x*(1-z)/piA^2)-
    sum(x*z*Q3^2/Q1)
  hes[1,2] <- hes[2,1] <- -sum((1-x)*z/(piA+piB-1)^2) - sum(x*z*Q2/Q1)
```

```
  hes[1,3] <- hes[3,1] <- -sum(x*z*piB*(muA1-y)*Q2/(Q1*s^2))
  hes[1,4] <- hes[4,1] <- 0
  hes[1,5] <- hes[5,1] <- sum(x*z*piB*(muB1-y)*Q2/(Q1*s^2))
  hes[1,6] <- hes[6,1] <- sum(x*z*piB*(muA1-muB1)*(muA1+muB1-2*y)*Q2/(Q1*s^3))
  hes[2,2] <- -sum((1-x)*z/(piA+piB-1)^2) - sum(x*z*Q4^2/Q1)
  hes[2,3] <- hes[3,2] <- sum(x*z*piA*(muA1-y)*Q2/(Q1*s^2))
  hes[2,4] <- hes[4,2] <- 0
  hes[2,5] <- hes[5,2] <- -sum(x*z*piA*(muB1-y)*Q2/(Q1*s^2))
  hes[2,6] <- hes[6,2] <- -sum(x*z*piA*(muA1-muB1)*(muA1+muB1-2*y)*Q2/(Q1*s^3))
  hes[3,3] <- -sum(x*z*Q3*piA*(Q3*piA*s^2+Q4*piB*(-muA1^2+s^2+2*muA1*y-y^2))/(Q1*s^4))
  hes[3,4] <- hes[4,3] <- 0
  hes[3,5] <- hes[5,3] <- -sum(x*z*piA*piB*(muA1-y)*(muB1-y)*Q2 / (Q1*s^4))
  hes[3,6] <- -sum(x*z*Q3*piA*(y-muA1)*
    (2*Q3*piA*s^2+Q4*piB*(muB1^2-muA1^2+2*s^2+2*muA1*y-2*muB1*y)) / (Q1*s^5))
  hes[6,3] <- hes[3,6]
  hes[4,4] <- -sum(x*(1-z)/s^2)
  hes[4,5] <- hes[5,4] <- 0
  hes[4,6] <- hes[6,4] <- sum(x*(1-z)*2*(muA0-y)/s^3)
  hes[5,5] <- -sum(x*z*piB*Q4*(Q4*piB*s^2 + Q3*piA*(s^2+2*muB1*y-muB1^2-y^2))/(Q1*s^4))
  hes[5,6] <- -sum(x*z*piB*(y-muB1)*Q4*(2*Q4*piB*s^2+Q3*piA*(muA1^2-muB1^2+2*s^2-2*muA1*y+2*muB1*y))/(Q1*s^5))
  hes[6,5] <- hes[5,6]
  hes[6,6] <- sum(x*(1-z)*(s^2-3*(muA0-y)^2)/(s^4))+
    sum(x*z*(Q3^2*piA^2*s^2*(s^2-3*(muA1-y)^2)+
    Q4^2*piB^2*s^2*(s^2-3*(muB1-y)^2)+
    Q2*piA*piB*(muA1^4 - 2*muA1^2*muB1^2 + muB1^4 - 3*muA1^2*s^2 - 3*muB1^2*s^2 +
    2*s^4 - 2*(muA1 + muB1)*(2*(muA1 - muB1)^2 - 3*s^2)*y +
    2*(2*(muA1 - muB1)^2 - 3*s^2)*y^2))/(Q1*s^6))

  hes
}
```

OA.3 Accuracy of the LS ATE

When the model is correctly specified, the LS ATE is consistent, like any other maximum-likelihood estimate. However, it may be biased in finite samples. In a simulation study, we find that this finite sample bias is minimal with a sample size of 100,000 (50,000 in treatment and 50,000 in control). For each set of parameter values in the simulation, we compute the average estimate of the ATE (averaged over the 2000 simulated data sets) and compare it to the true value to obtain an empirical estimate of the bias of $\widehat{\tau}^{\text{LS}}$. For the base parameter values, the mean estimate is 0.06221 versus a true ATE of 0.06200 (0.34% empirical bias). The left panel of Figure OA.3.1 plots the true ATE versus the mean of $\widehat{\tau}^{\text{LS}}$ for the 37 different parameter settings in the simulation study and shows that the bias is minimal across parameter settings. The right panel shows that $\widehat{\tau}^{\text{LS}}$ has modest positive bias relative to $\widehat{\tau}^{\text{DiM}}$. (Recall, DiM is is unbiased in finite samples, so any bias we

find there is due to sampling variation in the simulation.) Thus, we conclude that $\widehat{\tau}^{\text{LS}}$ is reasonably close to unbiased for our example application with sample size around 70,000 each in treatment and control.

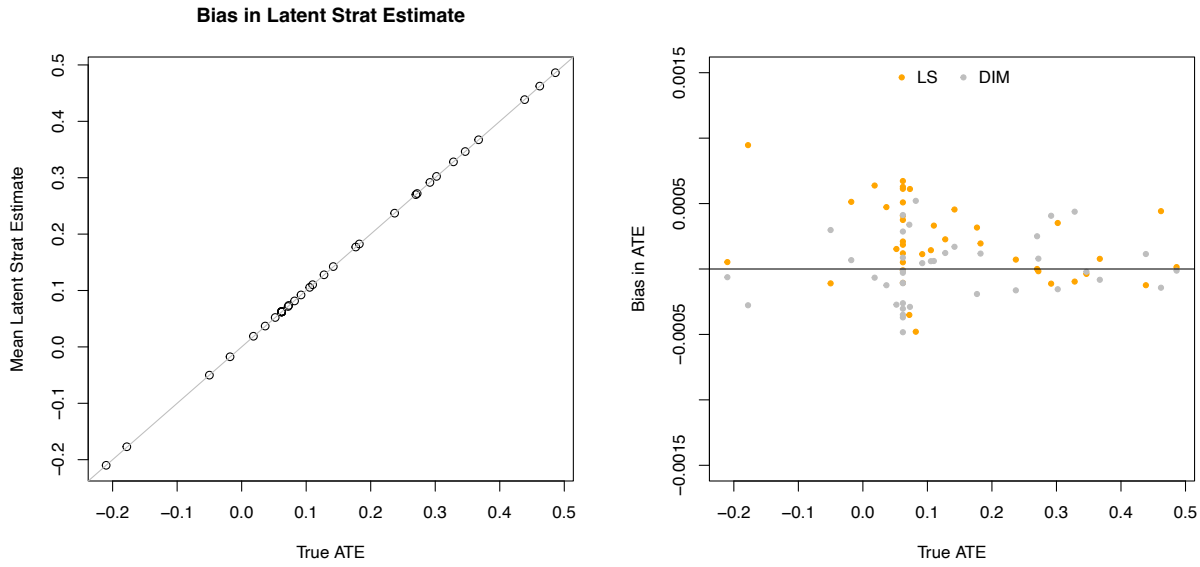

Figure OA.3.1: Simulation study shows empirical bias in $\widehat{\tau}^{\text{LS}}$ is minimal.

To confirm that the delta method produces a reliable estimate of the sampling variation of the ATE ($\widehat{\tau}^{\text{LS}}$), Figure OA.3.2 plots the empirical sampling variation from the simulation versus the average delta method estimate (across the 2000 simulated data sets). The plot shows that the delta method estimates are typically a little higher than the true sampling variation of the estimator. This suggests that the delta method provides a conservative estimate of the sampling variation in the MLE estimator, conditional on the model being correctly specified.

To evaluate the delta method standard errors with real data, we compare them to bootstrapped standard errors. Consistent with the simulation results in Figure OA.3.2, we find that standard errors computed by delta method are conservatively larger than bootstrap estimates for the catalog application. Table OA.3.2 reports the bootstrap standard errors which can be compared to the delta method standard errors in Table 3 in the paper.

When the model is misspecified, the LS ATE may be biased. Thus, it is important for users to consider whether the LS assumptions hold for their data and to use the misspecification test to detect misspecification. (See Section 2.4.) In this section, we quantify the bias in $\widehat{\tau}^{LS}$

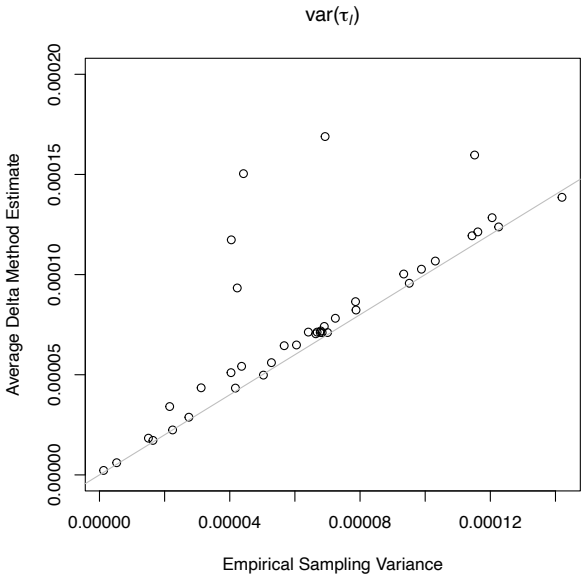

Figure OA.3.2: Empirical sampling variation of $\widehat{\tau}^{\mathrm{LS}}$ versus delta method estimate of sampling variation

Table OA.3.2: Comparison of ATE estimates for catalog experiments with standard errors computed by bootstrap

	Expt 1		Expt 2		Expt 3		Expt 4		Expt 5	
	est	se	est	se	est	se	est	se	est	se
$\widehat{\tau}^{\mathrm{DiM}}$	0.0040	0.0093	0.0247	0.0095	0.0245	0.0106	0.0248	0.0108	0.0110	0.0085
$\widehat{\tau}^{\mathrm{LS}}$	0.0189	0.0054	0.0230	0.0063	0.0243	0.0065	0.0290	0.0081	0.0190	0.0048

under three different potential misspecifications: (1) the outcomes are not normally distributed, (2) $\sigma_{B1} \neq \sigma_{A1} = \sigma_{A0}$, and (3) there is a fourth stratum that purchases under control, but not treatment. We also test the power of the IOS test to detect these misspecifications.

We extend the simulations to account for misspecification, by generating a synthetic data set with $N = 100,000$ customers from a model similar to LS, with parameters $\pi_A = 0.16$, $\pi_B = 0.01$, $\pi_C = 0.83$, $\mu_{A1} = 4.7$, $\mu_{A0} = 4.5$ and $\mu_{B1} = 3.0$. However, to create misspecification of the outcome distributions we add error terms drawn from a fatter-tailed t distribution with 3, 7 or 10 degrees of freedom or from a skewed shifted-Gamma distribution. The mean of the error terms is fixed to zero with standard deviation 1 (i.e. $\sigma_{A1} = \sigma_{A0} = \sigma_{B1} = 1$). We then compute $\widehat{\tau}^{\mathrm{LS}}$ and $\widehat{\tau}^{DiM}$ using

this data. We also computed the IOS test statistic and p-value using 100 bootstrapped samples from the simulated data set. We repeated this process with 100 simulated data sets, allowing us to estimate the bias of the LS estimator and the IOS test rejection rate at $p < 0.1$.

The results summarized in Table OA.3.3 show that the bias in $\widehat{\tau}^{\text{LS}}$ is rather modest. The magnitude of the estimated bias is 0.001 for both $\widehat{\tau}^{DiM}$ and $\widehat{\tau}^{\text{LS}}$, suggesting that the bias in $\widehat{\tau}^{\text{LS}}$ is within the sampling error of the simulation.[21] We also find that the IOS test fails to detect these misspecifications.

Table OA.3.3: Estimated bias in $\widehat{\tau}^{\text{LS}}$ when simulated error terms follow t or Gamma distributions.

Error Distribution	Bias in $\widehat{\tau}^{\text{LS}}$	Bias in $\widehat{\tau}^{DiM}$	IOS Rejection Rate	Mean IOS p-value
$t(3)$	-0.001	0.001	0.10	0.482
$t(7)$	0.001	0.001	0.12	0.463
$t(10)$	0.001	0.001	0.08	0.520
Gamma	0.000	0.001	0.10	0.537

We completed a similar simulation where the data was generated from Normal distributions, but with differing values of $\sigma_{B1} \neq \sigma_{A1} = \sigma_{A0} = 1$. Table OA.3.4 shows that $\widehat{\tau}^{\text{LS}}$ can be substantially biased when σ_{B1} is larger or smaller than $\sigma_{A1} = \sigma_{A0} = 1$. The IOS test reliably detects this misspecification when σ_{B1} is smaller than $\sigma_{A1} = \sigma_{A0} = 1$, but not when σ_{B1} is larger than $\sigma_{A1} = \sigma_{A0} = 1$.

Table OA.3.4: Estimated bias in $\widehat{\tau}^{\text{LS}}$ when $\sigma_{B1} \neq \sigma_{A1} = \sigma_{A0} = 1$.

σ_{B1}	Bias in $\widehat{\tau}^{\text{LS}}$	Bias in $\widehat{\tau}^{\text{DiM}}$	IOS Rejection Rate	Mean IOS p-value
0.75	0.012	-0.002	0.95	0.026
1.25	-0.015	-0.001	0.00	0.912

Finally, we repeated this analysis generating data from a model where there are four strata: the three strata in the LS model and a fourth stratum "D" where $Y(0) > 1$ and $Y(1) = 0$. We assumed that the response for the control group stratum D is Normal with mean $\mu_D = 4$ and standard deviation $\sigma_{D0} = \sigma_{A1} = \sigma_{A0} = \sigma_{B1}$. We varied the size of stratum D from $\pi_D = 0$ to 0.015, reducing the size of the A and B strata proportionally. As Table OA.3.5 shows, this misspecification produces substantial biases in $\widehat{\tau}^{\text{LS}}$. This misspecification is readily detected by the

[21] Since $\widehat{\tau}^{DiM}$ is unbiased, the estimated bias in $\widehat{\tau}^{DiM}$ represents sampling error.

IOS test when the D stratum is large ($\pi_D = 0.015$), however there is still substantial bias when $\pi_D = 0.005$ and this is only detected by the IOS test about half the time.

Table OA.3.5: Estimated bias in $\widehat{\tau}^{\text{LS}}$ when there is a fourth strata where $Y(0) > 1$ and $Y(1) = 0$. The size of this group is π_D.

π_D	Bias in $\widehat{\tau}^{\text{LS}}$	Bias in $\widehat{\tau}^{\text{DiM}}$	IOS Rejection Rate	Mean IOS p-value
0.000	0.002	-0.002	0.07	0.525
0.005	0.012	-0.001	0.46	0.226
0.010	0.027	-0.001	0.77	0.084
0.015	0.044	0.001	0.95	0.030

OA.4 Identification of two-component Normal mixture models

A general two-component Normal mixture model can be poorly-identified (Ho et al. 2022). However, as we show in this section, the LS model is much better identified because data observed in the control group serves to constrain the mixture model for the treated customers who buy. As we discuss in Section 2.2, the mixing proportion between the two components ($\pi_A/(\pi_A + \pi_B)$) is well-identified by the difference in purchase rates in treatment and control. Further, in our application, we assume that the variance of the Normal components is the same for A0, A1 and B0, and that common σ is well-identified by the observed variance for customers in control who purchase (A0). This additional structure in the LS model substantially improves the identification of the mixture.

To illustrate this, we fit an unconstrained two-component Normal mixture using the data for treated customers that buy in Experiment 2. That is, we fit the model with the likelihood

$$\ell(Y_i) = \pi \frac{1}{\sigma_{A1}} \phi\left(\frac{Y_i - \mu_{A1}}{\sigma_{A1}}\right) + (1-\pi)\frac{1}{\sigma_{B1}} \phi\left(\frac{Y_i - \mu_{B1}}{\sigma_{B1}}\right) \tag{17}$$

using the data that identifies the mixture model in LS. We then compare that to the LS parameter estimates using the full data set.

Table OA.4.6 shows parameter estimates for a general mixture and the LS model along with bootstrapped standard errors. The parameters for the unconstrained mixture model have large standard errors, particularly for the mixing ratio $\pi_A/(\pi_A+\pi_B)$[22] and the parameters of the smaller

[22]The standard mixing ratio in the two-component Normal mixture is $\pi_A/(\pi_A + \pi_B)$. We estimate this directly for the mixture model and indirectly for the LS model.

mixture component μ_{B1} and σ_{B1}. By contrast, the corresponding parameters in the LS model have standards errors that are an order of magnitude smaller. This illustrates that the LS model is substantially better identified than the unconstrained mixture model.

Table OA.4.6: Comparison of parameter estimates for an unconstrained two-component Normal mixture model versus the latent stratification model for Experiment 2 with bootstrap standard errors shown in parentheses.

	Unconstrained Mixture Model	Latent Stratification
$\pi_A/(\pi_A+\pi_B)$	0.870	0.978
	(0.243)	(0.008)
μ_{A1}	4.860	4.688
	(0.147)	(0.017)
μ_{B1}	3.258	2.992
	(0.691)	(0.439)
σ_{A1}	0.979	1.102
	(0.007)	(0.007)
σ_{B1}	0.894	
	(0.303)	
π_C		0.834
		(0.001)
μ_{A0}		4.616
		(0.011)
$\ell\ell$ treated, buy	-17371.6	-17398.2
N treated, buy	11,442	11,442
$\ell\ell$ all		-96265.8
N all		138,227

OA.5 Distribution of log-sales for the application in Section 3

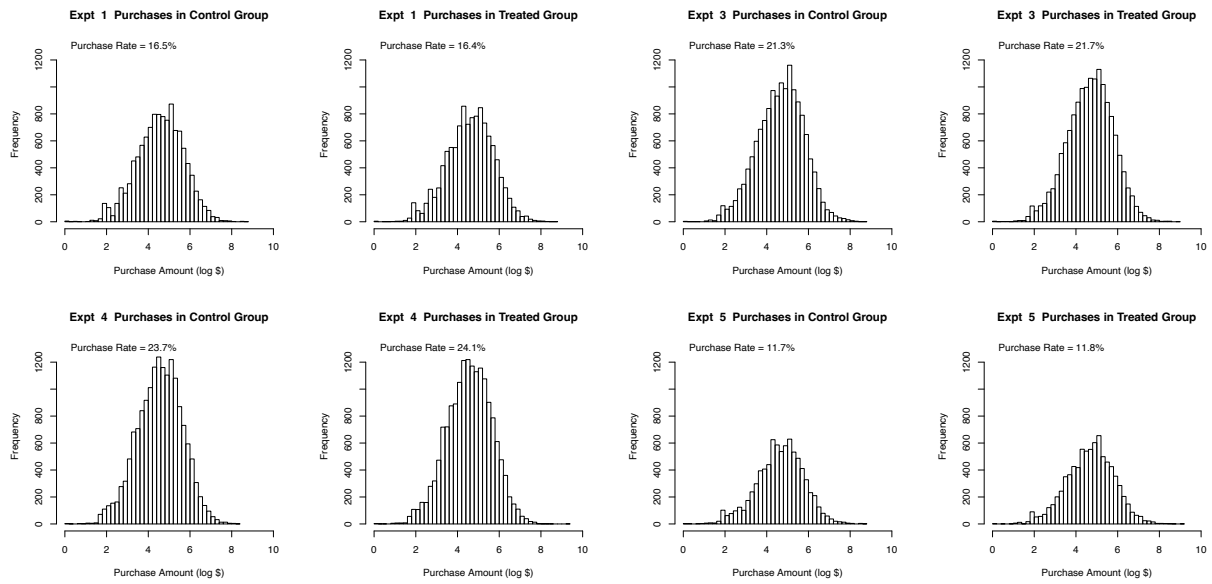

Figure OA.5.1: Distribution of log-purchase amounts for the application in Section 3

OA.6 Relaxing the constraint $\sigma_{A0} = \sigma_{A1} = \sigma_{B0}$

Table OA.6.1: Model estimates when the constraint $\sigma_{A0} = \sigma_{A1} = \sigma_{B0}$ is relaxed

	Expt 1		Expt 2		Expt 3		Expt 4		Expt 5	
	est	se	est	se	est	se	est	se	est	se
π_A	0.161	0.001	0.159	0.001	0.210	0.001	0.234	0.001	0.115	0.001
π_B	0.008	0.002	0.010	0.002	0.013	0.002	0.012	0.003	0.006	0.001
μ_{A0}	4.575	0.010	4.616	0.011	4.660	0.009	4.551	0.008	4.648	0.012
μ_{A1}	4.678	0.026	4.751	0.017	4.771	0.016	4.633	0.017	4.778	0.021
μ_{B1}	2.803	0.566	2.871	0.074	2.905	0.127	2.955	0.287	2.857	0.176
σ_{A1}	1.050	0.018	1.033	0.012	1.026	0.011	1.026	0.011	1.051	0.014
σ_{A0}	1.100	0.007	1.123	0.007	1.108	0.007	1.070	0.006	1.113	0.009
σ_{B1}	0.880	0.005	0.887	0.010	0.875	0.007	0.856	0.015	0.890	0.007
$\widehat{\tau}^{\mathrm{LS}}$	0.0346	0.0081	0.0468	0.0071	0.0533	0.0082	0.0473	0.0095	0.0309	0.0058

Table OA.6.1 provides parameter estimates for the catalog experiment with the $\sigma_{A0} = \sigma_{A1} = \sigma_{B0}$ constraint relaxed. The estimated values of σ_{A0}, σ_{A1}, and σ_{B0} are all similar to those in Table

2 in the paper, albeit the estimates of σ_{A1} and σ_{B0} and σ_{A0} slightly larger than the pooled variance.

OA.7 Alternative benefit of LS: reduced sample sizes

An alternative way to realize the benefit of latent stratification is to reduce the size of the total sample needed for an experiment to detect a specific increase in sales. The retailer that provided the data estimated that the cost of sending a catalog to customers was approximately one US dollar. An increase of one dollar on average in Experiment 2 between treatment and control translates to approximately a 5% increase in sales (and also a 5% increase in log-sales). Using a DiM analysis would have required approximately 72,000 consumers to detect such an effect, while using LS would have required approximately 38,000 consumers (almost half).

To compute the required sample size with equal allocation between treatment and control, we use the formulas:

$$n_1 = \left(1 + \frac{1}{\kappa}\right)(z_{\alpha/2} + z_{\beta})^2 \frac{\sigma^2}{d^2} \qquad n_0 = \kappa n_1$$

where $\alpha = 0.05$, $\beta = 0.2$, z_q is the upper q quantile of the standard Normal distribution, σ is the data's standard deviation, d is the effect size to be detected and κ is the ratio of control to treatment group sample sizes $\kappa = n_0/n_1$. The result is the required sample for each arm (treatment or control). Suppose we would like to detect a log-sales increase of 5% vs. the null hypothesis of no increase in log-sales. In Experiment 2 this increase would translate to an increase from log-sales of 0.747 to 0.784, or about \$1.06 in sales, which would cover the costs of sending the catalog to customers.

Using table 3, we can compute the standard deviation of the data under difference-in-means to be approximately $0.0095 \cdot \sqrt{70000/2} = 1.777$, while for latent stratification it is $0.0069 \cdot \sqrt{70000/2} = 1.29$. Plugging into the sample size formula, under difference-in-means we would need approximately 36k consumers per treatment, while under latent stratification we would need approximately 19k consumers per treatment.

Another way to realize the benefit is to use a substantially smaller control group. We can vary the ratio of sample sizes of the control and the treatment groups κ to make the control group as small as possible while leaving the total sample size fixed. For Experiment 2 with $n = 138,227$, in order to detect the same 5% increase in sales this implies that κ can be as low as 8.1%, yielding $n_0 = 10,307$ and $n_1 = 127,920$. If the true effect is an average increase in sales of \$1.19 (as it is

in the data for Experiment 2), and assuming a catalog cost of \$1, this approach could increase the firm's profit earned during the test by $(68914 - 10307) \cdot 0.19 = \$11,135$.

We can also look at the increase in power due to latent stratification given a fixed sample size of 140k consumers, equally allocated between treatment and control. The smaller standard error would have increased the power to detect a 3% increase in sales from 64% under difference-in-means to 89%.

OA.8 Regression adjustment with observed pre-randomization covariates

Table OA.8.2: Treatment effects with post-stratification using pre-rendomization covariates

	Experiment 40	Experiment 41	Experiment 42	Experiment 43	Experiment 48
Treatment	0.0044	0.0252**	0.0256**	0.0257**	0.0112
	(0.0086)	(0.0087)	(0.0097)	(0.0099)	(0.0078)
T	0.5431***	0.6148***	0.8229***	0.8493***	0.4806***
	(0.0349)	(0.0358)	(0.0396)	(0.0413)	(0.0320)
D	0.0007*	0.0009+	0.0005	0.0009+	0.0008+
	(0.0003)	(0.0005)	(0.0004)	(0.0005)	(0.0004)
I	-0.1803***	-0.0833*	0.0162	0.0285	-0.1234***
	(0.0365)	(0.0397)	(0.0411)	(0.0422)	(0.0367)
M	0.0018*	0.0014	0.0021*	0.0010	0.0016*
	(0.0007)	(0.0009)	(0.0008)	(0.0009)	(0.0008)
C	0.0019**	0.0025**	0.0013	0.0020*	0.0020*
	(0.0007)	(0.0009)	(0.0008)	(0.0009)	(0.0009)
F	3.1164***	2.9989***	3.4019***	3.3151***	2.1056***
	(0.0930)	(0.0948)	(0.0945)	(0.0883)	(0.0960)
R	0.0058***	0.0050**	-0.0052**	-0.0114***	0.0049**
	(0.0017)	(0.0017)	(0.0020)	(0.0020)	(0.0016)
Treatment × T	-0.0069	0.0150	0.0904	0.0875	-0.0127
	(0.0491)	(0.0504)	(0.0557)	(0.0580)	(0.0448)
Treatment × D	0.0000	-0.0009	-0.0008	-0.0010+	-0.0002
	(0.0005)	(0.0006)	(0.0005)	(0.0006)	(0.0005)
Treatment × I	0.0035	-0.0232	-0.0167	0.0573	0.0219
	(0.0509)	(0.0552)	(0.0567)	(0.0577)	(0.0503)
Treatment × M	-0.0004	0.0024*	0.0023*	0.0022+	0.0012
	(0.0010)	(0.0011)	(0.0011)	(0.0011)	(0.0010)
Treatment × C	0.0007	-0.0022+	-0.0016	-0.0018	-0.0008
	(0.0010)	(0.0012)	(0.0011)	(0.0011)	(0.0011)
Treatment × F	0.0060	0.0406	-0.1193	-0.1604	-0.1184
	(0.1256)	(0.1285)	(0.1299)	(0.1232)	(0.1270)
Treatment × R	-0.0016	0.0005	0.0018	-0.0011	-0.0013
	(0.0025)	(0.0025)	(0.0028)	(0.0028)	(0.0023)
Constant	-0.1580***	-0.1753***	-0.0179	0.1133***	-0.1555***
	(0.0226)	(0.0226)	(0.0251)	(0.0255)	(0.0207)
N	138,281	138,227	138,155	138,138	138,039
Adjusted R^2	0.182415	0.175661	0.181214	0.170583	0.130446

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Heteroskedasticity robust standard errors are in parentheses.

OA.9 Benefit of stratification with known strata memberships (oracle scenario)

We assume that the true data generating process is the one described by Equations (6), (7) and (8), with the main difference being that the stratum membership of each individual is observed, and derive the variances of $\widehat{\tau}^{\text{DiM}}$ and $\widehat{\tau}^{\text{LS}}$.

The difference-in-means estimator $\widehat{\tau}^{\text{DiM}} = \frac{\sum Y_i Z_i}{n_1} - \frac{\sum Y_i (1-Z_i)}{n_0}$ has a population variance of (Imbens and Rubin 2015, Eq. 6.17):

$$Var(\widehat{\tau}^{\text{DiM}}) = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}$$

where σ_0^2 is the population variance of $Y_i(0)$ and σ_1^2 is the population variance of $Y_i(1)$.

Under the LS model, we derive σ_1^2 and σ_0^2 from the mixture components as follows:

$$\begin{aligned}\sigma_1^2 &= \pi_A(\sigma^2 + \mu_{A1}^2) + \pi_B(\sigma^2 + \mu_{B1}^2) - \mu_1^2 \\ &= \pi_A(\sigma^2 + \mu_{A1}^2) + \pi_B(\sigma^2 + \mu_{B1}^2) - (\pi_A\mu_{A1} + \pi_B\mu_{B1})^2.\end{aligned}$$

The first equation comes from the properties of mixture distributions,[23] and the second from plugging in $\mu_1 = \pi_A\mu_{A1} + \pi_B\mu_{B1}$. Similarly,

$$\sigma_0^2 = \pi_A(\sigma^2 + \mu_{A0}^2) - \mu_0^2 = \pi_A(\sigma^2 + \mu_{A0}^2) - (\pi_A\mu_{A0})^2.$$

Summarizing, the expression for the variance is:

$$Var(\widehat{\tau}^{\text{DiM}}) = \frac{\pi_A(\sigma^2 + \mu_{A0}^2) - (\pi_A\mu_{A0})^2}{n_0} + \frac{\pi_A(\sigma^2 + \mu_{A1}^2) + \pi_B(\sigma^2 + \mu_{B1}^2) - (\pi_A\mu_{A1} + \pi_B\mu_{B1})^2}{n_1} \tag{18}$$

For the oracle scenario, the log-likelihood of the data equals:

$$\ell\ell = \sum_i \left(X_{iA}\log(\pi_A) + X_{iA}Z_i\log(f_{A1}(Y_i)) + \right. \tag{19}$$

$$+ X_{iB}\log(\pi_B) + X_{iB}Z_i\log(f_{B1}(Y_i)) + X_{iC}\log(1-\pi_A-\pi_B) \tag{20}$$

$$\left. + X_{iA}(1-Z_i)\log(f_{A0}(Y_i))\right) \tag{21}$$

where $f_{A1}(y)$ is the pdf of $\mathcal{N}(\mu_{A1}, \sigma^2)$, $f_{A0}(y)$ of $\mathcal{N}(\mu_{A0}, \sigma^2)$ and $f_{B1}(y)$ of $\mathcal{N}(\mu_{B1}, \sigma_{B1}^2)$, and where π_C was replaced by $1 - \pi_A - \pi_B$. Using the log-likelihood we can compute the standard error of the ATE using the inverse of the Fisher Information matrix and the delta method, as follows.

[23] See `https://en.wikipedia.org/wiki/Mixture_distribution`, section on Moments, Accessed December 8, 2021.

The gradient of the log-likelihood is:

$$\begin{aligned}
\frac{\partial \ell\ell}{\partial \pi_A} &= \sum_i \frac{X_{iA}}{\pi_A} - \sum_i \frac{X_{iC}}{1-\pi_A-\pi_B} \\
\frac{\partial \ell\ell}{\partial \pi_B} &= \sum_i \frac{X_{iB}}{\pi_B} - \sum_i \frac{X_{iC}}{1-\pi_A-\pi_B} \\
\frac{\partial \ell\ell}{\partial \mu_{A1}} &= \sum_i X_{iA} Z_i \frac{Y_i - \mu_{A1}}{\sigma^2} \\
\frac{\partial \ell\ell}{\partial \mu_{A0}} &= \sum_i X_{iA} (1-Z_i) \frac{Y_i - \mu_{A0}}{\sigma^2} \\
\frac{\partial \ell\ell}{\partial \mu_{B1}} &= \sum_i X_{iB} Z_i \frac{Y_i - \mu_{B1}}{\sigma_{B1}^2} \\
\frac{\partial \ell\ell}{\partial \sigma} &= \sum_i X_{iA} Z_i \frac{(Y_i-\mu_{A1})^2 - \sigma^2}{\sigma^3} + \sum_i X_{iA}(1-Z_i) \frac{(Y_i-\mu_{A0})^2 - \sigma^2}{\sigma^3} + \sum_i X_{iB} Z_i \frac{(Y_i-\mu_{B1})^2 - \sigma^2}{\sigma^3}
\end{aligned} \tag{22}$$

Denote $\theta = (\pi_A, \pi_B, \mu_{A1}, \mu_{A0}, \mu_{B1}, \sigma)^T$, then the Fisher information matrix for one observation $(X_{iA}, X_{iB}, Z_i, Y_i)$ is:

$$\begin{aligned}
I(\theta) &= -\mathbf{E}\left[\frac{\partial^2 LL}{\partial\theta\partial\theta'}\right] \\
&= \begin{pmatrix}
\frac{1}{\pi_A} - \frac{1}{\pi_A+\pi_B-1} & -\frac{1}{\pi_A+\pi_B-1} & 0 & 0 & 0 & 0 \\
-\frac{1}{\pi_A+\pi_B-1} & \frac{1}{\pi_B} - \frac{1}{\pi_A+\pi_B-1} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{p\pi_A}{\sigma^2} & 0 & 0 & 0 \\
0 & 0 & 0 & -\frac{(p-1)\pi_A}{\sigma^2} & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{p\pi_B}{\sigma^2} & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{2(p\pi_B+\pi_A)}{\sigma^2}
\end{pmatrix}
\end{aligned} \tag{23}$$

where $p = Pr(Z_i = 1)$. The Variance-Covariance matrix of θ is:

$$\begin{aligned}
V &= (nI(\theta))^{-1} \\
&= \begin{pmatrix}
\frac{(1-\pi_A)\pi_A}{n} & -\frac{\pi_A\pi_B}{n} & 0 & 0 & 0 & 0 \\
-\frac{\pi_A\pi_B}{n} & \frac{(1-\pi_B)\pi_B}{n} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{\sigma^2}{n_1\pi_A} & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{\sigma^2}{n_0\pi_A} & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{\sigma^2}{n_1\pi_B} & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{\sigma^2}{2n_1\pi_B+2n\pi_A}
\end{pmatrix}
\end{aligned} \tag{24}$$

The gradient of the LS estimator $\widehat{\tau}^{\text{LS}} = \widehat{\pi}_A(\widehat{\mu}_{A1} - \widehat{\mu}_{A0}) + \widehat{\pi}_B\widehat{\mu}_{B1}$ equals

$$g(\theta) = \frac{\partial g}{\partial \theta} = (\mu_{A1} - \mu_{A0}, \mu_{B1}, \pi_A, -\pi_A, \pi_B, 0)^T.$$

Then by the Delta method:

$$\begin{aligned}
&Var(\widehat{\tau}^{\text{LS}}) = g^T V g \\
&= \frac{\pi_A\sigma^2}{n_0} + \frac{\sigma^2(\pi_A + \pi_B)}{n_1} + \frac{\mu_{B1}\pi_B(2\mu_{A0}\pi_A - 2\mu_{A1}\pi_A + \mu_{B1}) - (\pi_A - 1)\pi_A(\mu_{A0} - \mu_{A1})^2 - \mu_{B1}^2\pi_B^2}{n} \\
&= \frac{\pi_A\sigma^2}{n_0} + \frac{\sigma^2(\pi_A + \pi_B)}{n_1} + \frac{\pi_B\mu_{B1}^2 + \pi_A(\mu_{A1} - \mu_{A0})^2 - (\pi_B\mu_{B1} + \pi_A(\mu_{A1} - \mu_{A0}))^2}{n} \qquad (25)
\end{aligned}$$

Summing up the terms together we get:

$$\begin{aligned}
Var(\widehat{\tau}^{\text{LS}}) = \pi_A\left(\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_0}\right) + (\mu_{A1} - \mu_{A0})^2\frac{\pi_A(1 - \pi_A)}{n} + \\
+ \pi_B\frac{\sigma^2}{n_1} + \mu_{B1}^2\frac{\pi_B(1 - \pi_B)}{n} - 2\frac{\pi_A\pi_B}{n}(\mu_{A1} - \mu_{A0})\mu_{B1} \qquad (26)
\end{aligned}$$

Finally, the difference between $Var(\widehat{\tau}^{\text{DiM}})$ and $Var(\widehat{\tau}^{\text{LS}})$ equals:

$$\begin{aligned}
Var(\widehat{\tau}^{\text{DiM}}) - Var(\widehat{\tau}^{\text{LS}}) = \frac{\pi_A\mu_{A0}^2 - (\pi_A\mu_{A0})^2}{n_0} + \frac{\pi_A\mu_{A1}^2 + \pi_B\mu_{B1}^2 - (\pi_A\mu_{A1} + \pi_B\mu_{B1})^2}{n_1} \\
- \left(\frac{\pi_A(\mu_{A1} - \mu_{A0})^2 + \pi_B\mu_{B1}^2 - (\pi_A(\mu_{A1} - \mu_{A0}) + \pi_B\mu_{B1})^2}{n}\right) \qquad (27)
\end{aligned}$$

We can verify that the difference is positive using the following Wolfram Mathematica code:

```
diff = (piA*muA0^2 - (piA*muA0)^2)/n0 +
(piA*muA1^2 + piB*muB1^2 - (piA*muA1 + piB*muB1)^2)/n1 -
(piA*(muA1 - muA0)^2 + piB*muB1^2 - (piA*(muA1 - muA0) +piB*muB1)^2)/n

Reduce[{diff > 0, piA > 0, piB > 0, piA + piB < 1,
muA0 > 0, muB1 > 0, muA1 > 0, n1 + n0 == n, n > 0}]
```

OA.10 Two-strata estimator

In this section we show that separating consumers into stratum C consumers, and combining the customers in the A and B strata into a single AB stratum, does not yield any variance reduction of the ATE. We divide the consumers into two strata: stratum C of consumers who don't buy regardless of treatment with size π_C, and another stratum of potential buyers with size $1 - \pi_C$,

in which consumers have outcomes with mean $\mu_{AB1} > 0$ under treatment and $\mu_{AB0} > 0$ under control.

We can estimate the size of the C stratum by counting the number of non-buyers in the treatment group as follows: $\widehat{\pi}_C = \frac{\sum_i \mathbb{I}(Y_i=0) Z_i}{n_1}$.

Using the estimated size of π_C we can estimate the ATE in the non-C stratum:

$$\widehat{\mu}_{AB1} = \frac{\sum_i Z_i Y_i}{(1-\widehat{\pi}_C) n_1} \tag{28}$$

$$\widehat{\mu}_{AB0} = \frac{\sum_i (1-Z_i) Y_i}{(1-\widehat{\pi}_C) n_0} \tag{29}$$

The stratified estimator of the ATE equals:

$$\widehat{\tau}^{\text{2S}} = (1-\widehat{\pi}_C)(\widehat{\mu}_{AB1} - \widehat{\mu}_{AB0}) + \widehat{\pi}_C \cdot 0 \tag{30}$$

$$= \frac{\sum_i Z_i Y_i}{n_1} - \frac{\sum_i (1-Z_i) Y_i}{n_0} = \widehat{\tau}^{\text{DiM}} \tag{31}$$

The last equation follows from plugging-in $\widehat{\mu}_{AB1}$ and $\widehat{\mu}_{AB0}$ and noticing that $(1-\widehat{\pi}_C)$ cancels out. The result is exactly the standard difference-in-means estimator.

OA.11 Empirical reduction in variance

This section reports the empirical sampling variation for τ^{LS} and τ^{DiM} over a range of parameter values. We begin with a base set of parameters similar to the estiamtes for Experiment 2 in the example application. We then vary these parameters one-at-a-time holding the other parameters fixed. For each set of parameter values, we generate 2000 data sets of size $n = 100,000$ from the LS model and estimate the ATE using both estimators. As a benchmark, we also estimate the sampling variance under the oracle scenario where the latent strata are observed, which provides a lower-bound on the sampling variation of the LS ATE.

Figure OA.11.2 shows the sampling variation of the ATE estimators as μ_{A1} and μ_{B1} are varied. These plots represent "slices" of Figure 4 in the main text, and show the sampling variation of τ^{LS} and τ^{DiM}, rather than the difference. Similar to Figure 4, Figure OA.11.2 shows that when μ_{A1} and μ_{B1} have similar values there is more sampling variation in $\widehat{\tau}^{\text{LS}}$; when μ_{A1} and μ_{B1} are close in value, the mixture model becomes weakly-identified. Specifically, in the left panel of Figure OA.11.2, when μ_{A1} is closer to 3 (the baseline value of μ_{B1}), the LS sampling variation is nearly the same as DiM and far from the oracle scenario. Similarly, in the right panel, LS actually has

slightly higher sampling variation than DiM when μ_{B1} is close to the baseline value of $\mu_{A1} = 4.7$. The benefit of latent stratification is greatest when μ_{A1} and μ_{B1} have different values.

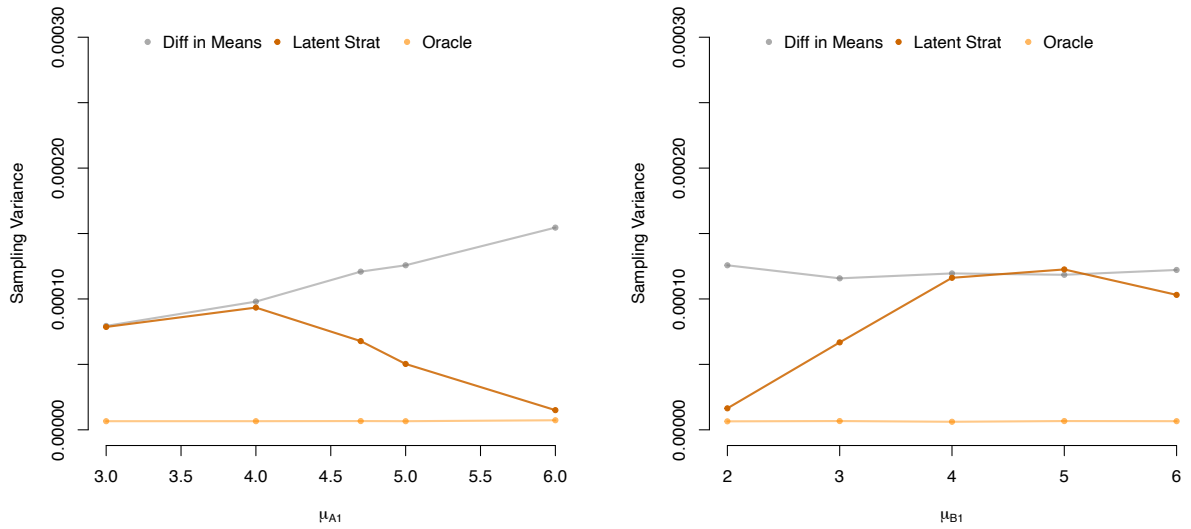

Figure OA.11.2: Sampling variation in alternative estimators of the ATE for different values of μ_{A1} and μ_{B1}. Other parameters are fixed at $\pi_A = 0.16$, $\pi_B = 0.01$, $\pi_C = 0.83$, $\mu_{A1} = 4.7$, $\mu_{A0} = 4.5$, $\mu_{B1} = 3$ and $\sigma_{A1} = \sigma_{A0} = \sigma_{B1} = 1$.

Figure OA.11.3 shows the sampling variation of the ATE estimators as μ_{A0} and σ are each varied. The left panel of Figure OA.11.3 shows that the LS estimator consistently has proportionally lower sampling variance than the DiM estimator across different values of μ_{A0}. The variance reduction $\left(1 - \frac{Var(\widehat{\tau}^{\text{LS}})}{Var(\widehat{\tau}^{\text{DiM}})}\right)$ does not vary substantially with μ_{A0}. However, the variance reduction is substantially affected by σ. The right panel of Figure OA.11.3 shows that when σ is low, $\widehat{\tau}^{\text{LS}}$ has nearly the same sampling variance as the oracle estimator. As σ increases, $Var(\widehat{\tau}^{\text{LS}})$ increases, due to the increased difficulty of estimating the mixture model in Equation 6. At $\sigma = 2.0$ the sampling variance of LS is nearly the same as DiM.

Figure OA.11.4 shows the sampling variation of τ^{LS} and τ^{DiM} as the strata proportions π_A, π_B and π_C are varied. We parameterize the simplex as $\frac{\pi_A}{\pi_A+\pi_B}$ and π_C. Figure OA.11.4 shows that the sampling variation of both estimators is quite sensitive to the mixing proportion. However, Figure OA.11.5 shows that the variance reduction $\left(1 - \frac{Var(\widehat{\tau}^{\text{LS}})}{Var(\widehat{\tau}^{\text{DiM}})}\right)$ does not vary substantially for intermediate values of $\frac{\pi_A}{\pi_A+\pi_B}$ and π_C. The reduction is highest when π_C is smaller (so there is

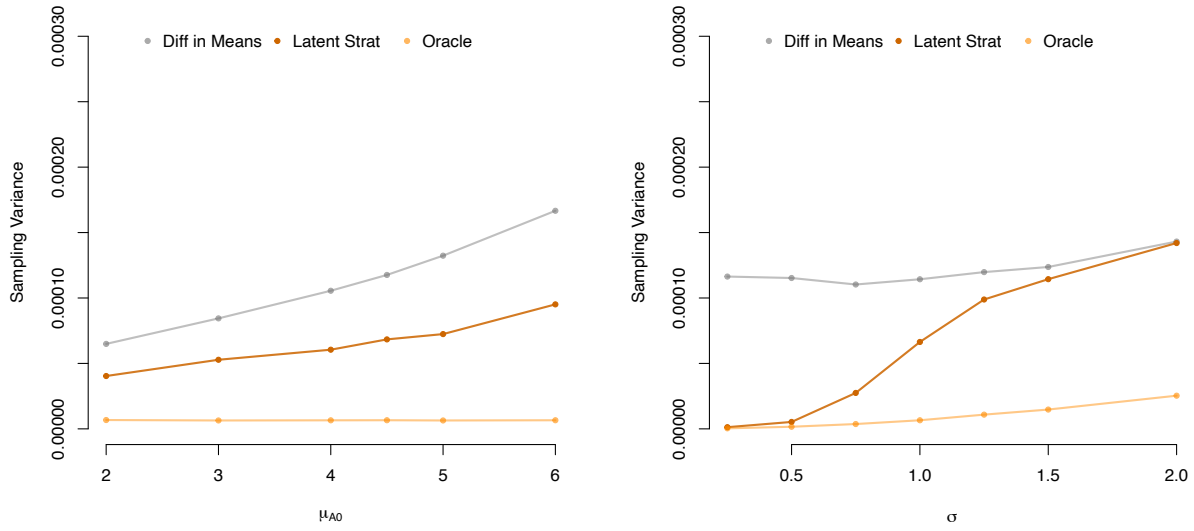

Figure OA.11.3: Sampling variation in alternative estimators of the ATE for different values of μ_{A0} and σ. Other parameters are fixed at $\pi_A = 0.16$, $\pi_B = 0.01$, $\pi_C = 0.83$, $\mu_{A1} = 4.7$, $\mu_{A0} = 4.5$, $\mu_{B1} = 3$ and $\sigma_{A1} = \sigma_{A0} = \sigma_{B1} = 1$.

more data available to estimate the mixture) and π_A is larger relative to π_B.

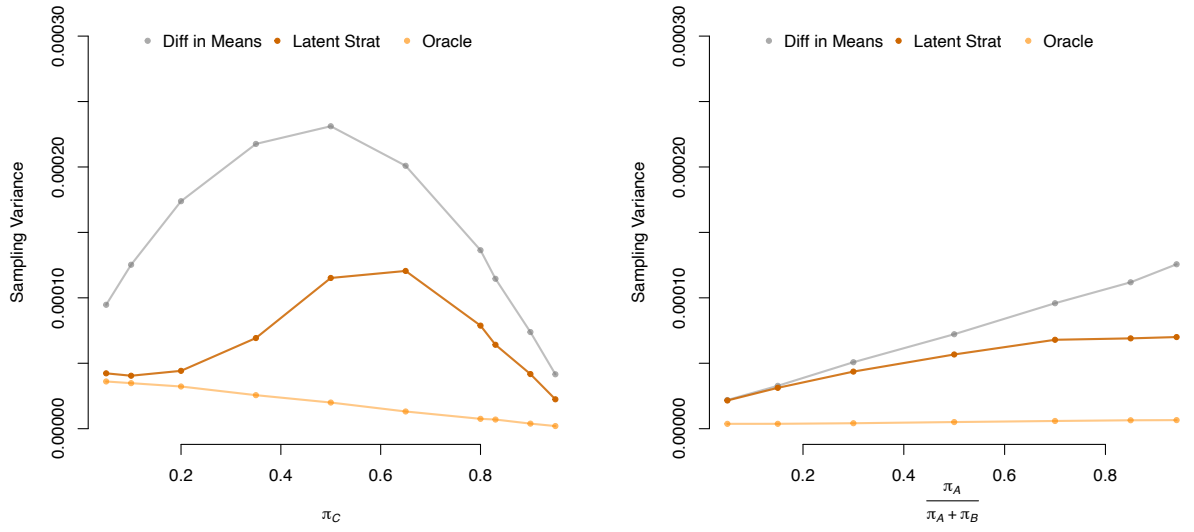

Figure OA.11.4: Sampling variation in alternative estimators of the ATE for different values of π_C and $\frac{\pi_A}{\pi_A+\pi_B}$. As π_C is varied, $\frac{\pi_A}{\pi_A+\pi_B}$ is fixed at 0.16/(0.16+0.01). As $\frac{\pi_A}{\pi_A+\pi_B}$ is varied π_C is fixed at 0.83. Other parameters are fixed at $\mu_{A1} = 4.7$, $\mu_{A0} = 4.5$, $\mu_{B1} = 3$ and $\sigma = 1$.

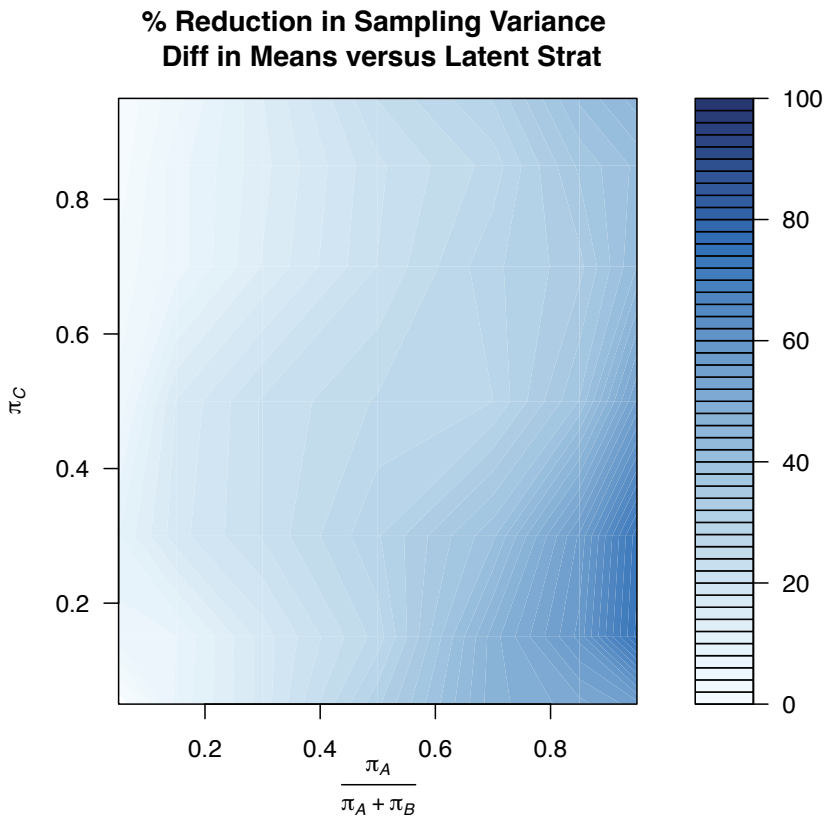

Figure OA.11.5: Percent reduction in sampling variation $\left(1 - \frac{Var(\widehat{\tau}^{\mathrm{LS}})}{Var(\widehat{\tau}^{\mathrm{DiM}})}\right)$ for different values of π_C and $\frac{\pi_A}{\pi_A+\pi_C}$. As π_C is varied, $\frac{\pi_A}{\pi_A+\pi_B}$ is fixed at 0.16/(0.16+0.01). As $\frac{\pi_A}{\pi_A+\pi_B}$ is varied π_C is fixed at 0.83. Other parameters are fixed at $\mu_{A1} = 4.7$, $\mu_{A0} = 4.5$, $\mu_{B1} = 3$ and $\sigma = 1$.