ESTIMATION OF TREATMENT EFFECT

Bob Agnew, raagnew1@gmail.com, www.raagnew.com

Introduction

This white paper distills what I have learned and applied over many years in marketing analytics. Most of the methods for estimating, and assessing the significance of, a marketing treatment involve random samples of individual customers or subjects, usually with a test group receiving a treatment and a control group not receiving it. The statistics are relatively simple, even when repeated measures on the same individuals are involved. Calculations are readily performed in SAS, R, other stat packages, or even in Excel. Only the last approach, based on aggregated time series, gets into more complex methodology.

Repeated Measures: Before and After

We have a sample of n individuals or subjects who are observed before and after some treatment with respect to a measure of interest; this can be continuous like sales, a count, or a binary $1 / 0$ response.

B_i = before-treatment measurement for individual $i = 1,...,n$

A_i = after-treatment measurement for individual $i = 1,...,n$

This is the classical repeated measures problem in statistics, which can be broader in scope (http://en.wikipedia.org/wiki/Repeated_measures_design). The estimated treatment effect is $\overline{A} - \overline{B}$, the difference of sample means. Significance can be assessed via a one-sample t-test on the differences $A_i - B_i$ to see if the estimated treatment effect is significantly different from zero. (PROC TTEST in SAS with no class variable; can also use the PAIRED statement in this simple setting.)

There are two problems with repeated measures. First, we may not have the luxury of repeated measures on the same subjects. Second, even if we do, there may be a time or trend effect which has nothing to do with the treatment. The main advantage is that variance is generally reduced when

observing the same subjects due to correlative behavior.

Controlled Experiment: Test vs Control

We have two independent samples during the *same* time period. The test group gets the treatment and the control group doesn't. It is assumed that the test and control groups are randomly sampled from the same population.

C_i = measurement for control group individual $i = 1,...,n$

T_i = measurement for test group individual $i = 1,...,m$

The estimated treatment effect is $\overline{T} - \overline{C}$, the difference of sample means. Significance can be assessed via a 2-sample t-test to see if the estimated treatment effect is significantly different from zero. (PROC TTEST in SAS with a class variable to distinguish the two groups.)

This analysis can also be performed via regression:

$Y_i = \alpha + \beta * D_i + \varepsilon_i$

Here we stack the test and control samples. Y_i is the measurement for individual i and D_i is a dummy indicator of test (1 for test, 0 for control). $\hat{\beta}$ is the estimated test effect and we can readily get the associated significance (p-value) from the regression output (e.g., from PROC REG in SAS). If the response is binary (e.g., response or activation), there is a tendency to want to employ logistic regression, but that isn't necessary and you lose the direct interpretability of the estimated effect. In the binary case, we have $E(Y_i) = p$ and $Var(Y_i) = p(1-p)$ for test, $E(Y_i) = q$ and $Var(Y_i) = q(1-q)$ for control. Normally, p and q aren't all that different, even when the test effect is significant, so the normal regression assumption of constant variance across the ε_i is reasonably close.

Actually, significance calculations in the binary case are extremely simple and they can be readily performed in Excel as follows.

$Var\left(\overline{T}-\overline{C}\right) = \frac{p(1-p)}{m}+\frac{q(1-q)}{n} \cong \frac{\overline{T}\left(1-\overline{T}\right)}{m}+\frac{\overline{C}\left(1-\overline{C}\right)}{n}$ so that

$Z = \frac{\overline{T}-\overline{C}}{\sqrt{\frac{\overline{T}\left(1-\overline{T}\right)}{m}+\frac{\overline{C}\left(1-\overline{C}\right)}{n}}}$ is approximately $N(0,1)$ for large samples. Hence,

for a one-sided test, 1 - $NORMSDIST(ABS(Z))$ represents estimated significance in terms of p-value. For the more conventional two-sided test, which makes no a priori assumption about test lift or suppression, $2*(1-NORMSDIST(ABS(Z)))$ is the estimated p-value.

In the case of counts, we can distinguish two cases at the individual level. If *transactions* during a time period consist of binary events (e.g., purchase of one unit), then we expect a Poisson distribution for the accumulation of individual counts in a time period and

$Var\left(\overline{T}-\overline{C}\right) \cong \frac{\overline{T}}{m}+\frac{\overline{C}}{n}$ so we can generate a p-value simply in Excel as before. If, on the other hand, transactions can be multiple, random amounts, we have a compound Poisson distribution for the accumulation of individual counts during a time period. In this case, assuming the same transaction parameters for test and control, we have

$Var\left(\overline{T}-\overline{C}\right) \cong \left(\frac{\overline{T}}{m}+\frac{\overline{C}}{n}\right)\left(\hat{v}+\frac{\hat{\xi}^2}{\hat{v}}\right)$ where v is the mean transaction quantity and ξ^2 is the variance of transaction quantity. Obviously, $v = 1$ and $\xi^2 = 0$ in the pure Poisson case. The reason for considering this, versus just performing a t-test or regression, is that v and ξ^2 can in principle be estimated from a single test or control sample and then applied thereafter, at least approximately, in other comparisons. The assumption here is that these behavioral parameters are homogeneous across individuals and have nothing to do with promotions and such. Unless we have confidence in this assumption, it's better to perform a regular t-test or regression. Moreover, this type of approximation gets more difficult when we combine test vs control with repeated measures.

Although test vs control is ubiquitous in direct marketing, it can be contaminated if the control sample is not from the same population as the test sample or if it receives extraneous treatments

not applied to the test group.

Combination: Test vs Control with Pre-Period Adjustment

Sometimes we are able to combine the previous two approaches, repeated measures and test vs control, to achieve significantly improved estimation efficiency. We now have four groups, before and after for test and control.

C_i^B = before-treatment measurement for control group individual $i = 1,...,n$

C_i^A = after-treatment measurement for control group individual $i = 1,...,n$
(but control group gets no treatment)

T_i^B = before-treatment measurement for test group individual $i = 1,...,m$

T_i^A = after-treatment measurement for test group individual $i = 1,...,m$
(only test group gets a treatment)

The estimated treatment effect is

$$\left(\overline{T}^A - \overline{T}^B\right) - \left(\overline{C}^A - \overline{C}^B\right) = \left(\overline{T}^A - \overline{C}^A\right) - \left(\overline{T}^B - \overline{C}^B\right)$$

$\overline{T}^B - \overline{C}^B$ is sometimes called the pre-post adjustment to the test vs control comparison, but the meat is in the left-hand expression where we usually benefit from repeated measures variance reduction. Significance of the estimated treatment effect can be assessed using a two-sample t-test with observations $T_i^A - T_i^B$ in one sample and $C_i^A - C_i^B$ in the other. (PROC TTEST with a class variable to distinguish the two groups.)

Once again, this analysis can also be performed via regression:

$$Y_i = \alpha + \beta * D_i + \varepsilon_i$$

Here we stack the test and control difference samples, i.e., measurement Y_i is $T_i^A - T_i^B$ for a test observation and $C_i^A - C_i^B$ for a control observation while D_i is a dummy indicator of test (1 for test, 0 for control). $\hat{\beta}$ is the estimated test effect and its associated p-value is available in the regression output (e.g., from PROC REG in SAS).

This combined approach, sometimes called pre-post adjustment, has the same limitations as before. We may not have the luxury of repeated measures within the test and control groups. Moreover, the control group can be contaminated relative to the test group. Even if repeated measures are feasible, it is possible that estimation efficiency will not be improved by including pre-period results. It usually improves because of correlation across the periods for the same individuals, but it is at least conceivable that there is no (or even negative) benefit and that a simple test versus control analysis is superior. The next approach generalizes to allow flexible weighting on the pre-period.

ANCOVA: Flexible Pre-Period Weighting

ANCOVA stands for Analysis of Covariance. It is a general procedure for bringing in regression covariates to better isolate a test effect. The setup is the same as the last section. However, here we run the following regression:

$$A_i = \alpha + \beta * B_i + \gamma * D_i + \varepsilon_i$$

The test and control samples are again stacked with A_i after test, B_i before test, and D_i a dummy indicator of test (1 for test, 0 for control). As before, only test individuals receive a treatment. $\hat{\gamma}$ is the estimated test effect and $\hat{\beta}$ is a flexible weight on the pre-period results for both test and control. Once again, we have all the usual regression outputs available for significance assessment. Usually $\hat{\beta}$ is between zero and one and significant, but it may be less than one (default in the previous section). In my experience, $\hat{\beta}$ is often in the 0.7 neighborhood.

All the usual regression caveats apply here. First, we assume that the test and control samples are randomly drawn from the same population. We assume equal variances across test and control (i.e., across the ε_i). We are also dismissing any other relevant covariates that might sharpen the fit. But ANCOVA in this form does get past the issue of whether to do a pre-post adjustment or not by providing a flexible weight on the pre-period. Moreover, within the confines of the model, this version of ANCOVA provides an optimal result in terms of estimation efficiency.

Difference in Differences

Difference in differences (DID) is increasingly used in the econometrics community to estimate a treatment effect where we have test versus control and before and after, but we don't necessarily have repeated measures on the same individuals and the test and control groups aren't necessarily so cleanly defined (http://en.wikipedia.org/wiki/Difference-in-differences). For example, the test and control groups might consist of individuals living in two disjoint sets of DMAs, obviously not randomly drawn from the same population, and there could be different individuals in the before and after samples.

Once again we can employ a regression model:

$$Y_i = \alpha + \beta * TEST_i + \gamma * AFTER_i + \delta * TEST_i * AFTER_i + \varepsilon_i$$

Here all observations have been stacked with dummies for demarcations.

Y_i = measurement or outcome variable

$TEST_i$ = 1 for test group, 0 for control group

$AFTER_i$ = 1 for after treatment, 0 for before

As an approximation, we are assuming independent observations here, even when we have a significant number of repeated observations on the same individuals before and after. This will tend to understate significance. $\hat{\beta}$ reflects a general test vs control effect, aside from treatment, assumed constant across before and after. $\hat{\gamma}$ reflects a general after vs before "trend" effect, assumed constant across test and control. $\hat{\delta}$ reflects the treatment effect for test combined with after (only the test group gets treatment).

Once again, the usual regression assumptions apply, like constant variance across the ε_i . Nevertheless, we get all the usual regression outputs and significance estimates. And this is a pretty general technique, allowing estimation of a treatment effect across somewhat disparate test and control groups, subject of course to reasonability constraints.

It turns out that the estimates here are particularly simple with $\hat{\alpha} = \overline{C}^B$, $\hat{\beta} = \overline{T}^B - \overline{C}^B$, $\hat{\gamma} = \overline{C}^A - \overline{C}^B$, and $\hat{\delta} = \left(\overline{T}^A - \overline{C}^A\right) - \left(\overline{T}^B - \overline{C}^B\right)$. Note that $\hat{\delta}$ is the same estimated test effect as we got in pre-post adjustment, however the significance estimate will be different because

we aren't explicitly accounting for repeated measures. In fact, we might be able to approximate $Var\left(\hat{\delta}\right)$ and develop an associated Z-score and p-value without regression, as indicated previously, but in most instances it will be more straightforward and expeditious to just perform the regression to get a p-value.

One caveat. When we actually have a clean test vs control experiment with before and after repeated measures on the same individuals, then the estimated test effect from DID will be identical. However, the significance calculation will not account for repeated measure correlation. It's best to consider this DID approach only when we don't have a pure experimental setting.

Time Series Modeling

This approach utilizes aggregated observations over time so we are no longer utilizing observations at the individual customer level. We have a test group that receives a treatment over a time period subsequent to a presumed business-as-usual (BAU) period of sufficient length to employ time-series methods. We then build a BAU model and forecast into the test period, comparing actuals to forecasts to estimate the incremental test effects. Most time-series models also provide confidence bands around the forecasts for future observations which can be employed to assess significance relative to specified confidence levels. Measurements are usually counts, sales, or ratios of such to the customer base. We prefer modeling ratios, or "intensities," and converting estimates back to metrics of interest using actual customer counts. Obviously, actual customer counts would not be known in pure forecasting, but this is not a pure forecasting application. This is, rather, a hybrid application.

If there is no control group, we rely on time trend and seasonals in our BAU model. This can be done via simple regression with linear time trend, seasonal dummies, and perhaps epochal dummies. Moreover, there is a wide variety of univariate methods available, including ARIMA and generalized smoothing. Suffice to say that we have utilized PROC ARIMA in SAS for the former and PROC ESM in SAS for the latter (with either the SEASONAL or ADDWINTERS option).

What I choose to focus on here is the case where we do have a control, which presumably is at least approximately comparable to test. Hence, in our BAU model (assuming time periods are months)

we have the following time-series regression:

$T_i = \alpha + \beta * C_i + \varepsilon_i$

over prior time periods $i = 1,...,n$.

T_i = test measurement in month i

C_i = control measurement in month i

Here $\hat{\alpha}$ is the estimated intercept and $\hat{\beta}$ is the estimated control coefficient. We don't include any additional linear time trend beyond control because it can be distortive. Other epochal drivers, like special program dummies, can be carefully included if relevant. The error term ε_i may be autocorrelated and seasonal. Our preferred method accounts for and models both. The forecast for future month $m+n$ is thus $\hat{T}_{m+n} = \hat{\alpha} + \hat{\beta} * C_{m+n} + \hat{\varepsilon}_{m+n}$ and $T_{m+n} - \hat{T}_{m+n}$ is the estimated incremental test effect where $\hat{\varepsilon}_{m+n}$ is explicitly modeled to reflect both autocorrelation and seasonality.

For this type of modeling, we have employed PROC AUTOREG in SAS. In this procedure, you specify lags for autocorrelation correction; if you allow all lags ≤ 12 and use the BACKSTEP option, the procedure will determine significant lags for you (e.g., SLSTAY = .10 or default = .05). Lag 12 accounts for monthly seasonality (beyond control) so there is no need for seasonal dummies in the model. PROC AUTOREG produces forecasts and also confidence bands around those forecasts with a specified level of significance (e.g., ALPHA = .15). In the test period, actuals can be compared to forecasts and also to confidence limits. Moreover, forecasts, and, at least approximately, confidence bands can be accumulated across multiple months in the test period. In general, PROC AUTOREG is straightforward to use with good diagnostics and automated lag selection. It accounts for residual autocorrelation and seasonality, but it does not account for residual moving average effects. That would require ARIMA transfer function estimation.

Short of full transfer function modeling, we have done univariate modeling of test-control intensity differences, again using PROC ARIMA and PROC ESM in SAS, so here we are forcing a unit coefficient on control. These methods don't work well with less than 36 months of data, and they

require judgmental manipulation, but they are certainly contenders to the AUTOREG approach.

On occasion, we have employed a much cruder time-series forecasting methodology, which doesn't account for trend, seasonality, or autocorrelation, only randomness, and is best applied to test-control intensity differences, if at all. Let

B_i = test intensity - control intensity difference in beforehand BAU month $i = 1,...,n$

Assume these observations are independent and identically distributed with sample mean $\overline{B}$ and sample standard deviation S. Let $\overline{A}$ be the sample mean in the test period $i = m+1,...,m+n$. Then, $\overline{A}-\overline{B}$ is the estimated test effect and

$$Z = \frac{\overline{A}-\overline{B}}{\sqrt{\frac{1}{m}+\frac{1}{n}}\sqrt{\frac{nS^2}{n-1}}}$$

is approximately t-distributed with $n-1$ degrees of freedom under the null hypothesis of no change in the underlying mean. We assume here a common standard deviation across the BAU and test periods although we have chosen to estimate randomness across the BAU period alone, similar to forecast models. Obviously, one could in principle run a normal t-test with presumed independent samples of n and m observations and with assumed common variance. In any event, this methodology is admittedly crude. It can be useful, however, in cases where trend, seasonality, and autocorrelation are minor issues and where historical BAU months are limited, or where you are simply seeking a rough benchmark. It is also easy to implement in Excel utilizing AVERAGE, STDEV or VAR, and $TDIST(ABS(Z), n-1, 2)$ to produce a 2-sided p-value for significance.

One issue associated with any time-series method is the number of *relevant* past months available. Some gurus specify a minimum of 50 months, but we are often short of that. We have scraped by with as few as 30 months with AUTOREG. Another issue, as before, is potential contamination of the control group. If the control group receives extraneous treatments in the test period that have not touched the test group, then estimated incrementals will be distorted. The bottom line is that the test and control groups don't have to be precisely comparable in terms of composition, but confounding control treatments should be

avoided if possible.

Additional Comments

If test and control aren't randomly sampled from the same population, you have to be very careful about *absolute* comparisons in the test period. The reason is obvious. Test may have been riding consistently above or below control historically, in the BAU, so an absolute gap can't be interpreted as significant, even if statistics look significant. In other words, the observed test vs control difference may not be due to the test treatment at all. Even when test and control are randomly sampled from the same population, we can often benefit from before and after modeling when repeated measures are involved.

Another issue is choice of technique. When should you use individualized models versus aggregated time series? With individualized models, you have to carry large datasets into your analysis, but you can get by with a shorter BAU period. With time-series, except for generalized smoothing, you are normally assuming that all historical months are equally important, but that may not be the case in that older history could be largely irrelevant. On the other hand, focusing on just a recent historical BAU period could be myopic in that there could have been unusual vagrant promotions affecting one or both test and control samples. The best situation, of course, is when these various techniques provide similar inferences about test lift or suppression, but that isn't always the case. When there are major differences, we have to delve into the causes of those differences.

Don Hedeker has provided good insights in the past on these estimation topics, particularly pre-post adjustment and ANCOVA. Don is Professor of Biostatistics, School of Public Health Sciences, University of Chicago (health.bsd.uchicago.edu/People/3129). Any errors or omissions in this white paper are, of course, my responsibility.