On balance tests

Let Y_i be a continuous, or discrete, variable and T_i a dummy variable which takes either values one or zero with probability p and $1-p$, respectively. A regression of Y_i on T_i and a constant is equivalent to comparing means for the two groups, one with values of $T=0$ and the other with values of $T=1$. It can be shown that:

$$\beta = \frac{cov[Y,T]}{var(T)} = \frac{p(1-p)(E[Y_i|T_i=1] - E[Y_i|T_i=0])}{p(1-p)} = E[Y_i|T_i=1] - E[Y_i|T_i=0]$$

In case of endogenous regressor, mean comparison and linear regression suffer from the same bias. To see this:

$$\begin{aligned} E[Y_i|T_i=1] = & \quad \alpha + \beta + E[\varepsilon_i|T_i=1] \\ E[Y_i|T_i=0] = & \quad \alpha + E[\varepsilon_i|T_i=0] \\ E[Y_i|T_i=1] - E[Y_i|T_i=0] = & \quad \beta + \underbrace{E[\varepsilon_i|T_i=1] - E[\varepsilon_i|T_i=0]}_{\text{selection bias}} \end{aligned}$$

A numerical example is shown in Table 1 with replication material on GitHub.

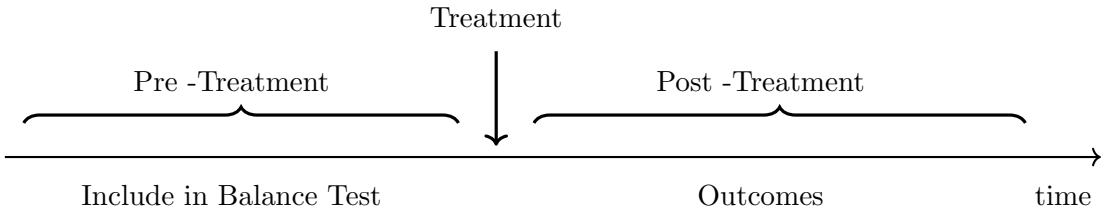
Table 1: Comparisons of means

	(1)	(2)	(3)
T	0.699	0.740	0.696
	(0.013)	(0.024)	(0.020)
Mean-Difference	0.699	0.740	0.696
Standard Error	0.013	0.024	0.020

Notes: The relationship between Y and T is $Y = 0.5 + 0.7 \cdot T + \varepsilon$ in column (1), $Y = 0.5 + 0.7 \cdot T + 0.3 * X_1 + \varepsilon$ in column (2) and $Y = 0.5 + 0.7 \cdot T + 0.15 * X_2 + \varepsilon$ where X_2 is itself a linear function of X_1. The coefficient estimate and standard error in parenthesis, shown in the top part of the table are from a regression of Y on T. In the bottom part of the table the mean difference between the two groups, those with $T=1$ and $T=0$, as well as the corresponding standard error is shown.

A covariate balance test helps us understand if treatment and control group are actually comparable. The variables to include in a balancing test are these variables which are usually measured pre-treatment. See Figure 1 for a visual representation. Geographical variables such as latitude, longitude, altitude, ruggedness, distance, surface are also unlikely to be affected by the treatment and can be included in a balance test.

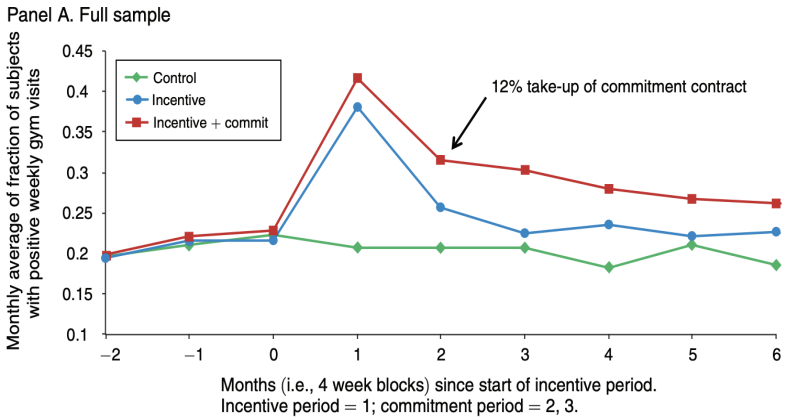
Figure 1: Timeline treatment



If the variables are measured post-treatment, it's likely they are affected by the treatment, therefore they should not be included in a balancing test as they are outcomes.

We encourage you to use visual representations to show covariate balance if possible. A nice example of pre and post mean comparison is shown in Figure 2 where raw means are plotted before and after treatment. An alternative way would be using coefficient plots.

Figure 2: Incentives, commitments, and habit formation



Notes: Figure from Royer, Heather, Mark Stehr, and Justin Sydnor. "Incentives, commitments, and habit formation in exercise: evidence from a field experiment with workers at a fortune-500 company." American Economic Journal: Applied Economics 7.3 (2015): 51-84.

In many cases, it makes sense to have a mean comparison controlling for other factors. For example, if the assignment is stratified, you would want to look the regression coefficient once you have controlled for other stuff i.e include school fixed, η_s, effects because the treatment is stratified at the school level. ρ is the parameter of interest.

$$x_i = \alpha + \rho T_i + \eta_s + \varepsilon_i$$

If you are testing if one, or several covariates, are affecting treatment, straightforward tests are bivariate regressions of the form:

$$T_i = \alpha + \beta x_i + \varepsilon_i$$

The β coefficients inform us if one or some covariates affect treatment individually.

In case you want to check whether several variables affect the treatment jointly, the specification to run is:

$$T_i = \alpha + X_i^{\intercal}\gamma + \eta_s + \varepsilon_i$$

Note that X_i is a vector of covariates. An F-test on the joint significance, whether all covariates jointly affect treatment, should be reported.