1. We ran logit regressions for questions 43521 and 450. Both models converged and consistently yielded p-values greater than 0.10 for nearly every dummy variable included in the mode, across all combinations of tariff and stimulus. This suggests that there is no compelling evidence of imbalance in the data, and that the treatment groups in this experiment were as good as randomly assigned for the observed responses to these two particular survey questions.

What about regression A1? It has two variables that are significant (kwh\_2009\_07 and kwh\_2009\_12). Why do you suppose we can ignore (or not ignore) that evidence of imbalance?

1. Including all available survey variables in the regression is essentially over-fitting the model. While the model may perfectly explain the results of this particular experiment, it would lack the flexibility to accurately predict outcomes in different scenarios and would thus be devoid of any external validity. In general, the benefit of understanding every single variable’s role in the observed results is outweighed by the model’s inability to adapt to new scenarios. Additionally, including multiple variables that control for the same underlying characteristic (e.g. income) is unnecessary and could lead to multicollinearity. Finally, using all available survey data could increase the number of observations that are dropped from the regression if some variables include numerous missing values.
2. Using only a subset of the available survey questions would avoid the trap of an over-fitted model described above. If a model based on this limited subset seemed to have a sufficient goodness-of-fit and enough explanatory power, this would confer advantages both in terms of heightened external validity and reduced complexity and computational overhead.

Good write up!