45/45

Evaluation of Section 2, Part A

1. The groups do appear balanced. While there were three significant variables in the logit model, two of these become insignificant in the t-test. There were two significant variables in the t-test, but "the presence of a few isolated statistically significant differences is usually also attributable to chance."[1] None of the consumption variables are significant in either test, so past consumption does not influence the likelihood of being treated.

2. A logit model is an easy way to check for imbalance, because you can include all the variables and check the significance level of each controlling for all other variables. However, it can also be problematic due to multicollinearity or perfect separation. Multicollinearity, when variables are highly correlated with each other, can cause standard errors to blow up. This causes variables to appear insignificant in the logit model when they are in fact significant (leading to a conclusion of balance when in fact the groups are imbalanced). If there is perfect separation – when an independent variable perfect predicts assignment to a treatment or control group – a coefficient cannot be estimated in the logit model because there is no variation across outcomes.

3. Adding more dummy variables to the logit model without adding more observations increases the chance of perfect separation or multicollinearity. The more variable categories you have, for example, the greater the risk of sparse dummy variables. The solution to this problem is to remove rows of data. However, this can cause perfect separation and can lead to too little data to estimate the model. In addition, the more variables you have, the greater the chance that some are highly correlated with each other. This multicollinearity problem can lead to drawing incorrect conclusions from the logit model.

Evaluation of Section 2, Part B

1. Given that the results in Part A showed the data was not imbalanced, propensity score weighting would not be necessary.

2. The coefficient estimate of the treatment-trial interaction variable is –0.0274 in the model without weights and –0.0236 in the model with weights. There is not much difference between the two models. If there had been imbalance in the data, the two models would have estimated much different coefficients. This supports the answer in question 1.

3. Because the data is not biased, the policy analyst would choose the fixed effects model without weights.

4. The coefficient on the treatment-trial interaction term in the fixed effects model without weights is –0.0274, meaning there was a 2.74% decrease in energy consumption for the B4 treatment group relative to the control group. This coefficient has a p-value of 0.058 so it is significant at the 10% level.

Randomization

1. The goal in evaluating the effectiveness of a policy or program is to measure the treatment effect on the outcome of interest. Specifically, we are interested in the average treatment effect: $ATT = E[Y_i(1) - Y_i(0) \mid D_i=1]$. However, the treatment effect of each individual is not possible to compute due to the fact that we observe only one outcome – treated or not treated. Simply comparing the outcomes for individuals that are treated and individuals that are not treated results in the expression $E[Y_i \mid D_i=1] - E[Y_i \mid D_i=0] = E[Y_i(1) \mid D_i=1] - E[Y_i(0) \mid D_i=1] + E[Y_i(0) \mid D_i=1] - E[Y_i(0) \mid D_i=0] = ATT + E[Y_i(0) \mid D_i=1] - E[Y_i(0) \mid D_i=0] = ATT + \textit{Selection Bias}$. When there is no random assignment, the baseline outcome for individuals who are treated is not the same as the baseline for individuals who are not treated. This difference is referred to as selection bias and will bias the estimated treatment effect of the policy or program. Random assignment mitigates the problem of selection bias. Randomly assigned treatment and control groups come from the same underlying population and therefore have the same conditional expectations, $E[Y_i(0) \mid D_i=1] = E[Y_i(0) \mid D_i=0]$ *Selection Bias = 0.*

2. One example of a corrupted randomization stage is the Western Massachusetts Electric Company experiment. Instead of households receiving a random assignment into treatment or control groups, households were filtered out and each treatment group was chosen sequentially. In the first stage, select households were removed (call this group F_1) and the remaining households were randomly divided into the first treatment group (T_1) and a control group. Of this control group, another set of households was removed (F_2) and the remaining households were randomly divided into the second treatment group (T_2) and a control. This process was repeated, creating a third filtered group (F_3), a third treatment group (T_3) and a control group. This remaining control group, as well as F_1, F_2, and F_3, became the overall control group for the trial. However, because select households were filtered out at each stage, the treatment groups didn't look anything like the control group. A second example of a corrupted randomization stage occurred in the CUB Energy Saver Program in Illinois. For this program, individuals sign up on their own. Instead of randomly being assigned to a treatment or control group, this program was opt-in. Not only did most participants live in Chicago, making the treatment group different from the population of Illinois as a whole, but there was also no control group. Data for other people in Illinois could be obtained to use as a control group, but this may be practically difficult as it is expensive.

3. One reason why participants might switch from their original assignment group is that the participant receives less technology than assigned. For example, a participant assigned to use an in-home display may never receive the unit. Another reason assignments may change is if individuals were randomized before it was determined if the household was even eligible for the program (may be assigned to a treatment they cannot receive).

4. The difficulty with a citywide energy policy is that it cannot be applied to a random treatment group. The policy is in effect for the entire city, so this is not a randomized controlled trial. However, a different city that did not implement a new energy policy could be used as a control group. To control for differences between the two cities, propensity score or covariate matching / weighting can be used to properly estimate the effect of this policy change. The propensity score measures the probability of participating in the program. If additional data exists for individuals or households in the target city and another city, covariates or the propensity score can be used to control for factors that may have influenced assignment (differences that may have affected implementation of this policy in the first place, or underlying population differences between the two cities) and remove the bias.

Big Data

1. (1) Time and energy required to clean data and create a useable format.
   (2) Storage of big data is a challenge, and datasets cannot be downloaded raw.
   (3) Generating quality data (free of bias) can be challenging, but generating "garbage" data is easy. You must be aware of the data generating process when working with big data.

2. "P-hacking" is the use of data to create results that look statistically significant, even if they are not. The more scenarios are tested, the more likely it is to find one that produces significant results. This process is made easier by Big Data. As the number of observations increases, uncertainty decreases, so everything becomes significant. This is problematic because factors that seem important (e.g. policies that seem effective) may not be in reality.

3. Over-fitting occurs when models are too complicated (too many explanatory variables are used to describe an outcome). Over-fit models may describe one particular set of observed data well, but they perform horribly when doing prediction. One way to prevent over-fitting is to introduce a penalty for models that are too complicated. This process should produce a model that is in between trivial / linear and too complicated. This method uses a penalized log-likelihood of the form $l^*(\beta) = l(\beta) - \lambda J(\theta)$, where $J(\theta)$ is a penalty function and λ is a tuning parameter. One possible penalty function is the L_1-penalization: $J(\theta) = \Sigma_i \|\beta_i\|$. This penalty function penalizes for all non-zero coefficients – the smaller the model, the smaller the penalty. To obtain the optimal tuning parameter, you can plot the cross-validated likelihood (using a portion of your data to build a model, using the model to predict the remaining data points, and comparing the predictions with the actual data to create a model with the best performance) against λ and choose the λ that gives the maximum log-likelihood.

4. With a classification algorithm to select households for participation in the new government program, the specific factors used to select households are known. Knowing these factors exactly can help the policy makers select an equivalent control group or calculate propensity scores to correct for the discrepancies and determine the effectiveness of the program. However, this process requires judgments on the policy makers' part on which households might benefit most from the program. A voluntary opt-in program would allow those households most enthusiastic about the program to participate, and these households would most likely see the largest benefit. Of course, in this process there are unobservables affecting a household's decision to opt-in, and it would be harder to determine the true effectiveness of the program.

---

1. Joshua D. Angrist and Jörn-Steffen Pischke, *Mastering 'Metrics* (Princeton and Oxford: Princeton University Press, 2015), 21. ↩