**Group Task 3: Section II Answers**

**PLEASE BE MORE CONCISE! -1 point**

**6.1**

We selected questions 200, 310, and 450. Q310 is for the employment status of the chief income earner of the household and Q450 is for the architecture structure of the house/apartment. House type could be a proxy for socioeconomic status, would affect the level of electricity usage, and could also affect the extent to which a household could make energy-saving upgrades to appliances or the home structure.

In the logit model, Y=1 indicates treatment and Y=0 indicates the control group. The significance level (p-value) of each independent variable indicates whether the monthly energy consumption, gender, employment status, or housing type predicts treatment status. For the purpose of this assignment, we used significance level p<0.05 to analyze the results.

After checking each logit model, we found only Tariff “B”, Stimulus “1” is completely balanced, which means none of the independent variables is able to predict treatment or control status. For Tariff “B,” Stimulus “3,” D310\_3 (self-employed respondents) is significant and its coefficient is positive; these respondents were more likely to be assigned to treatment. Tariff “A”, Stimulus “1” is the most imbalanced model, kwh\_2009\_07 and kwh\_2009\_12 are statistically significant. We would want to check the models using a larger sample of the data to help address the perfect separation we encountered with some models as well as to confirm the imbalances found.

While there is evidence of imbalance, so few variables for any given model are ever significant. How do you know that the imbalances are not due to chance alone? If a very large fraction of variables were significant, then one should begin to worry.

Furthermore, based on your results, A3 is also “completely imbalanced”.

**6.2**

The benefits of including all the variables are:

* Minimizing the chance that you miss an attribute that is unbalanced at baseline, because you’re accounting for many, many different participant attributes.

The disadvantages of including all the variables are:

* A reduced number of observations are included in the analysis, because households with missing data for any variable will be dropped. This leads to a undesirable smaller sample size, and potential bias if there is something different about respondents who do not answer certain questions.
* If using a small sample size, adding all of the variables increases the chances for variables to be dropped due to perfect separation, as well as the likelihood that the regression will fail.
* Using many independent variables increases the chance of variables being correlated with each other. The redundant questions would cause colinearity in the logit model.

**6.3**

Using a subset of questions rather than everything in the survey is more appropriate if independent variables may be redundant--and therefore pose issues with multicollinearity. In constructing the subset, you would want to choose variables which are likely highly correlated with electricity consumption, and therefore will be most important for balance in the final analysis of treatment effect. In general, a smaller subset of questions is also appropriate when you only have a few observations for some of the other variables in the dataset, and may need to be dropped from analysis.

Therefore, choosing a smaller subset with important independent variables allows you to check for balance while retaining enough observations in the sample size, as well as avoiding multicollinearity.

**Regression Results**

-----------------------------------------------------------------

**LOGIT where Treatment is Tariff = A, Stimulus = 1**

-----------------------------------------------------------------

-----------------------------------------------------------------

Perfect Separation produced by ['D\_310\_7']. Removed.

-----------------------------------------------------------------

Optimization terminated successfully.

Current function value: 0.599639

Iterations 6

Logit Regression Results

==============================================================================

Dep. Variable: T No. Observations: 367

Model: Logit Df Residuals: 349

Method: MLE Df Model: 17

Date: Tue, 07 Apr 2015 Pseudo R-squ.: 0.04824

Time: 16:59:48 Log-Likelihood: -220.07

converged: True LL-Null: -231.22

LLR p-value: 0.1731

===============================================================================

coef std err z P>|z| [95.0% Conf. Int.]

-------------------------------------------------------------------------------

const -0.8537 0.971 -0.879 0.379 -2.757 1.049

kwh\_2009\_07 0.0066 0.003 2.272 0.023 0.001 0.012

kwh\_2009\_08 -0.0035 0.002 -1.579 0.114 -0.008 0.001

kwh\_2009\_09 -0.0016 0.003 -0.554 0.580 -0.007 0.004

kwh\_2009\_10 0.0002 0.003 0.054 0.957 -0.005 0.006

kwh\_2009\_11 -0.0025 0.002 -1.008 0.314 -0.007 0.002

kwh\_2009\_12 0.0029 0.001 2.104 0.035 0.000 0.006

D\_200\_2 0.4382 0.237 1.851 0.064 -0.026 0.902

D\_310\_2 -0.0160 0.494 -0.032 0.974 -0.985 0.953

D\_310\_3 0.3281 0.471 0.697 0.486 -0.594 1.250

D\_310\_4 -1.4096 1.092 -1.291 0.197 -3.549 0.730

D\_310\_5 -0.1285 0.653 -0.197 0.844 -1.408 1.151

D\_310\_6 -0.4020 0.268 -1.500 0.134 -0.927 0.123

D\_450\_2 0.2495 0.964 0.259 0.796 -1.640 2.139

D\_450\_3 -0.1668 0.974 -0.171 0.864 -2.077 1.743

D\_450\_4 -0.2589 0.989 -0.262 0.793 -2.197 1.679

D\_450\_5 -0.0410 0.963 -0.043 0.966 -1.929 1.847

D\_450\_6 0.4523 1.741 0.260 0.795 -2.960 3.864

===============================================================================

-----------------------------------------------------------------

**LOGIT where Treatment is Tariff = A, Stimulus = 3**

-----------------------------------------------------------------

-----------------------------------------------------------------

Perfect Separation produced by ['D\_450\_6']. Removed.

-----------------------------------------------------------------

Optimization terminated successfully.

Current function value: 0.596821

Iterations 5

Logit Regression Results

==============================================================================

Dep. Variable: T No. Observations: 363

Model: Logit Df Residuals: 345

Method: MLE Df Model: 17

Date: Tue, 07 Apr 2015 Pseudo R-squ.: 0.03416

Time: 16:59:48 Log-Likelihood: -216.65

converged: True LL-Null: -224.31

LLR p-value: 0.5722

===============================================================================

coef std err z P>|z| [95.0% Conf. Int.]

-------------------------------------------------------------------------------

const -0.2385 0.843 -0.283 0.777 -1.891 1.414

kwh\_2009\_07 0.0022 0.003 0.797 0.425 -0.003 0.008

kwh\_2009\_08 -0.0024 0.002 -1.177 0.239 -0.006 0.002

kwh\_2009\_09 0.0015 0.003 0.532 0.594 -0.004 0.007

kwh\_2009\_10 0.0012 0.003 0.398 0.691 -0.005 0.007

kwh\_2009\_11 0.0015 0.002 0.632 0.527 -0.003 0.006

kwh\_2009\_12 -0.0023 0.002 -1.544 0.123 -0.005 0.001

D\_200\_2 -0.0224 0.238 -0.094 0.925 -0.490 0.445

D\_310\_2 0.6585 0.463 1.422 0.155 -0.249 1.566

D\_310\_3 0.5607 0.465 1.205 0.228 -0.351 1.473

D\_310\_4 0.9518 0.544 1.750 0.080 -0.114 2.018

D\_310\_5 -0.2311 0.720 -0.321 0.748 -1.643 1.181

D\_310\_6 -0.2487 0.282 -0.881 0.378 -0.802 0.305

D\_310\_7 -0.5893 1.140 -0.517 0.605 -2.823 1.645

D\_450\_2 -0.7099 0.838 -0.847 0.397 -2.352 0.932

D\_450\_3 -0.5228 0.848 -0.617 0.537 -2.184 1.139

D\_450\_4 -0.7480 0.861 -0.869 0.385 -2.436 0.940

D\_450\_5 -0.7951 0.845 -0.941 0.347 -2.452 0.862

===============================================================================

-----------------------------------------------------------------

**LOGIT where Treatment is Tariff = B, Stimulus = 1**

-----------------------------------------------------------------

-----------------------------------------------------------------

Perfect Separation produced by ['D\_310\_5', 'D\_450\_6']. Removed.

-----------------------------------------------------------------

Optimization terminated successfully.

Current function value: 0.405015

Iterations 6

Logit Regression Results

==============================================================================

Dep. Variable: T No. Observations: 286

Model: Logit Df Residuals: 269

Method: MLE Df Model: 16

Date: Tue, 07 Apr 2015 Pseudo R-squ.: 0.04323

Time: 16:59:48 Log-Likelihood: -115.83

converged: True LL-Null: -121.07

LLR p-value: 0.8411

===============================================================================

coef std err z P>|z| [95.0% Conf. Int.]

-------------------------------------------------------------------------------

const -1.0672 1.238 -0.862 0.389 -3.494 1.360

kwh\_2009\_07 -0.0005 0.004 -0.124 0.902 -0.008 0.007

kwh\_2009\_08 -0.0020 0.003 -0.669 0.504 -0.008 0.004

kwh\_2009\_09 0.0015 0.003 0.471 0.637 -0.005 0.008

kwh\_2009\_10 0.0059 0.004 1.535 0.125 -0.002 0.013

kwh\_2009\_11 -0.0032 0.003 -0.926 0.355 -0.010 0.004

kwh\_2009\_12 -0.0012 0.002 -0.565 0.572 -0.005 0.003

D\_200\_2 -0.1326 0.351 -0.378 0.705 -0.820 0.555

D\_310\_2 -0.1888 0.786 -0.240 0.810 -1.729 1.352

D\_310\_3 -0.2039 0.816 -0.250 0.803 -1.803 1.396

D\_310\_4 0.4845 0.862 0.562 0.574 -1.206 2.175

D\_310\_6 -0.2596 0.406 -0.640 0.522 -1.055 0.536

D\_310\_7 0.4496 1.174 0.383 0.702 -1.850 2.750

D\_450\_2 -0.6103 1.231 -0.496 0.620 -3.024 1.803

D\_450\_3 -0.5563 1.253 -0.444 0.657 -3.012 1.899

D\_450\_4 -1.4116 1.323 -1.067 0.286 -4.005 1.182

D\_450\_5 -0.6807 1.236 -0.551 0.582 -3.104 1.742

===============================================================================

-----------------------------------------------------------------

**LOGIT where Treatment is Tariff = B, Stimulus = 3**

-----------------------------------------------------------------

-----------------------------------------------------------------

Perfect Separation produced by ['D\_310\_7', 'D\_450\_6']. Removed.

-----------------------------------------------------------------

Perfect Separation produced by ['D\_310\_7', 'D\_450\_6']. Removed.

-----------------------------------------------------------------

Optimization terminated successfully.

Current function value: 0.364274

Iterations 7

Logit Regression Results

==============================================================================

Dep. Variable: T No. Observations: 285

Model: Logit Df Residuals: 268

Method: MLE Df Model: 16

Date: Tue, 07 Apr 2015 Pseudo R-squ.: 0.07233

Time: 16:59:48 Log-Likelihood: -103.82

converged: True LL-Null: -111.91

LLR p-value: 0.4399

===============================================================================

coef std err z P>|z| [95.0% Conf. Int.]

-------------------------------------------------------------------------------

const -2.3545 1.381 -1.705 0.088 -5.061 0.352

kwh\_2009\_07 -0.0016 0.004 -0.398 0.691 -0.009 0.006

kwh\_2009\_08 -0.0017 0.003 -0.551 0.582 -0.008 0.004

kwh\_2009\_09 -0.0017 0.004 -0.386 0.700 -0.010 0.007

kwh\_2009\_10 -0.0003 0.005 -0.074 0.941 -0.009 0.009

kwh\_2009\_11 0.0046 0.004 1.165 0.244 -0.003 0.012

kwh\_2009\_12 -0.0003 0.002 -0.132 0.895 -0.004 0.004

D\_200\_2 0.1503 0.372 0.404 0.686 -0.579 0.880

D\_310\_2 -0.9791 1.118 -0.876 0.381 -3.171 1.212

D\_310\_3 1.4136 0.603 2.343 0.019 0.231 2.596

D\_310\_4 0.9362 0.762 1.228 0.219 -0.558 2.430

D\_310\_5 -0.3952 1.132 -0.349 0.727 -2.615 1.824

D\_310\_6 -0.2737 0.446 -0.614 0.539 -1.147 0.600

D\_450\_2 0.2878 1.369 0.210 0.834 -2.395 2.971

D\_450\_3 0.2079 1.375 0.151 0.880 -2.486 2.902

D\_450\_4 0.6522 1.352 0.482 0.629 -1.997 3.302

D\_450\_5 -0.4268 1.387 -0.308 0.758 -3.145 2.292

===============================================================================