# Appendix A
# Analysis Details

# Regression Model Details

The generalized model equations with customer specific intercepts are shown in Equations 1 and 2 below, for both the gas and electric models.

**EQUATION 1: GAS MODEL REGRESSION EQUATION**

$$C_{it} = \alpha_i + \tau_t + \beta_b x_{b,it} P_{it} + \beta_h x_{h,it} P_{it} HDD_{it} + \gamma_{i,h} HDD_{it} + \varepsilon_{it}$$

**EQUATION 2: ELECTRIC MODEL REGRESSION EQUATION**

$$C_{it} = \alpha_i + \tau_t + \beta_b x_{b,it} P_{it} + \beta_h x_{h,it} P_{it} HDD_{it} + \beta_c x_{c,it} P_{it} CDD_{it} + \gamma_{i,h} HDD_{it} + \gamma_{i,c} CDD_{it} + \varepsilon_{it}$$

*Where*

$C_{it}$ is the monthly consumption for the household *i* in period *t*, expressed in kWh or Ccf per day,

$\alpha_i$ is the "customer-specific" intercept for household *i*, accounting for unexplained difference in use between households associated with the number of occupants, appliance holdings, lifestyle, etc.,

$\tau_t$ is the "time-specific" error for period t, reflecting the unexplained difference in use between time periods,

$x_{b,it}$ and $x_{h,it}$ are the dummy variables indicating the base measures (water heating, refrigeration), heating measures (envelope, heating system repair/replacement) and cooling measures (air-conditioning, etc.) that were installed at household *i* (1 if measure was installed, 0 if not)

$P_{it}$ is the dummy variable to designate the post-period for home *i* in time period *t* (0 in the pre-period and 1 in the post-period),

$\beta_b$, $\beta_h$ and $\beta_c$ are the regression estimators for the base, heating measures and cooling measures, representing the Ccf or kWh saved from base measures, Ccf or kWh per HDD for heating measures, and Ccf or kWh per CDD for cooling measures

$\gamma_{i,h}$ and $\gamma_{i,c}$ are the heating and cooling slopes for home *i*

$HDD_{it}$ and $CDD_{it}$ are the heating and cooling degree days for household *i* in period *t*

$\varepsilon_{it}$ is the error term that accounts for the difference between the model estimate and actual consumption for household *i* in period *t*.

# Regression Output

The regression results for the final gas and electric models are presented in the tables below. The $R^2$ for the gas model was 0.96 and 0.81 for the electric model.

**TABLE A- 1: GAS REGRESSION OUTPUT**

| Parameter | Estimator[1] | t-value[2] | Unit of Estimator | Homes in Model |
|---|---|---|---|---|
| Hot Water | -0.0738 | -6.21 | Ccf/Day | 1,149 |
| Insulation | -0.0274 | -21.32 | Ccf/HDD | 434 |
| Air Sealing | -0.0109 | -21.73 | Ccf/HDD | 2,933 |
| Duct Sealing | -0.0013 | -9.80 | Ccf/HDD | 493 |
| Heating Equipment | -0.0100 | -5.11 | Ccf/HDD | 106 |
| Multiple Measures[4] | -.03818 | -3.36 | Ccf/HDD | 19 |
| Multiple Measures[4] | -.02162 | -29.15 | Ccf/HDD | 1,330 |
| Heating Slope[5] | 0.1505 | 936.68 | Ccf/HDD | 5,862 |
| Intercept[5] | 0.0950 | 15.07 | Ccf/Day | 5,862 |

[1] For heating and cooling measures, the estimator represents the average change in use per degree day for homes in that measure group. For base use measures (Hot Water Conservation), it represents the average change in use per day.

[2] The t-value measures whether the value of the coefficient is statistically different from zero. It is calculated as the estimator divided by its standard error. A t-value with an absolute value of 1.64 or higher indicates the coefficient is statistically different from zero at the 90% confidence level.

[3] Single family homes.

[4] Homes with too many different measures to be placed into any measure group. These groups were differentiated into homes with heating system measures and those that did not have heating system measures.

[5] Represents the average for all homes in model. The full regression output includes a heating slope and intercept estimate for each individual home in the model.

**TABLE A- 2: ELECTRIC REGRESSION OUTPUT**

| Parameter | Estimator[1] | t-value[2] | Unit of Estimator | Homes in Model |
|---|---|---|---|---|
| Lighting | -1.0051 | -26.51 | kWh/Day | 13,584 |
| Hot Water Conservation | -1.2959 | -16.06 | kWh/Day | 1,981 |
| Refrigerator | -1.865 | -19.35 | kWh/Day | 1,041 |
| Insulation Heating | -0.3464 | -15.13 | kWh/HDD | 322 |
| Insulation Cooling[4] | -0.7147 | -6.27 | kWh/CDD | 277 |
| Air Sealing Heating | -0.1340 | -13.98 | kWh/HDD | 1,582 |
| Air Sealing Cooling[4] | -0.5851 | -12.09 | kWh/CDD | 1,384 |
| Duct Sealing Heating | -0.2191 | -17.09 | kWh/HDD | 648 |
| Duct Sealing Cooling[4] | -0.5132 | -12.45 | kWh/CDD | 1,610 |
| Heat Pumps | -0.3702 | -9.35 | kWh/HDD | 99 |
| Other Base Measures[5] | -0.7475 | -3.70 | kWh/Day | 174 |
| Multiple Heating Measures[6] | -0.2159 | -9.11 | kWh/HDD | 173 |
| Heating Slope[7] | 1.2077 | 163.59 | kWh/HDD | 23,201 |
| Cooling Slope[7] | 2.0769 | 211.79 | kWh/CDD | 23,201 |
| Intercept | 25.6076 | 134.88 | kWh/Day | 23,201 |

[1] For heating and cooling measures, the estimator represents the average change in use per degree day for homes in that measure group. For base use measures, such as Hot Water Conservation, it represents the average change in use per day.
[2] The t-value measures whether the value of the coefficient is statistically different from zero. It is calculated as the estimator value divided by its standard error. A t-value with an absolute value of 1.64 or higher indicates the coefficient is statistically different from zero at the 90% confidence level.
[3] Population encompasses single family homes only.
[4] The cooling and heating portions of the envelope measures were combined to estimate the final measure group savings.
[5] Group includes appliances and other unidentifiable measures.
[6] Homes with too many different measures to be placed into any measure group.
[7] Represents the average for all homes in model. The full regression output includes a heating/cooling slope and intercept estimate for each individual home in the model.

**TABLE A- 3: HOMES IN SINGLE FAMILY PROGRAM POPULATION BY MEASURE TYPE**

| | Number of Homes in Model | | Number of Homes in Program | |
|---|---|---|---|---|
| **Measure Type** | **Natural Gas** | **Electric** | **Natural Gas** | **Electricity** |
| Hot Water Cons. | 1,149 | 1,981 | 6,328 | 8,167 |
| Insulation | 434 | 599 | 2,293 | 3,760 |
| Air Sealing | 2,933 | 2,966 | 6,549 | 22,313 |
| Duct Sealing | 493 | 2,258 | 928 | 5,441 |
| Heating Equipment | 106 | 0 | 168 | 4 |
| Lighting | NA | 13,584 | NA | 38,088 |
| Refrigerator | NA | 1,041 | NA | 1,370 |
| Heat Pumps | NA | 99 | NA | 266 |
| Miscellaneous[1] | 0 | 0 | 182 | 534 |
| **Total Homes** | 5,862 | 23,201 | 8,298 | 39,932 |

[1] Includes Windows and Doors, Thermostats, Natural Gas Water Heaters, Electric Appliances, and Heat Pump Water Heaters

## Regression Diagnostics

After the savings were estimated, several diagnostics were checked for additional information. The model was tested for the following violations of assumptions:

- Autocorrelation (observations are not completely independent)
- Influential data points
- Heteroscedasticity (unequal variances)

Autocorrelation of errors is most common in time-series due to the intrinsic relationship between the most recent prior period and the present measurement, while unspecified variables are missing that would explain the underlying mechanisms for these changes. If the model exhibits autocorrelation, the estimators are unbiased, but the variance in the model tends to be artificially low.

Influential data points could occur when a small number of homes have a substantial upward or downward impact on the result.

If a data set exhibits heteroscedasticity, the estimator should be unbiased, but the variance is larger than may actually be the case. However, a high-degree of heteroscedasticity may also be a sign that the model is misspecified, which could affect the results.

In addition, multicollinearity occurs when predictor variables are correlated with one another, which can happen if measures are installed as a group. If multicollinearity is present, the estimators are sometimes of the wrong sign or not statistically significant.

The table below outlines the additional analyses used to verify the results and the findings we obtained.

**TABLE A- 4: DIAGNOSTICS OVERVIEW**

| Step | Result | | Finding | Implications |
|---|---|---|---|---|
| | Gas | Electric | | |
| Durbin-Watson statistic | 0.85 | 2.09 | This statistic indicates the presence of autocorrelation. | Autocorrelation is common in billing models; the estimators are unbiased but the variance may be understated. |
| Influential Data Points | 3 Identified | None | Homes with potential influential impacts were identified using a pooled DFFITS by household.[1] | The natural gas model was tested with and without the potential influential homes; these three homes had little impact on the final results (less than 1% change in the RR). |
| Goldfeld-Quandt test | 3.32 | 2.88 | This indicator measures heteroscedasticity in the data set. | A GQ statistic of 1.0 indicates no heteroscedasticity; 3.32 suggests a low level of heteroscedasticity that would not be expected to impact the results. |

[1] Belsley, D.A., Kuh, E., and Welsch, R.E. Regression Diagnostics, New York: John Wiley & Sons, Inc., 1980.