LBNL-59041

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Development of a Mathematical Air-Leakage Model from Measured Data

Jennifer McWilliams and Melanie Jung

January 2006

This work was also supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Program, of the U.S. Department of Energy under contract No. DE-AC03-76SF00098.

Development of a Mathematical Air-Leakage Model from Measured Data

Jennifer McWilliams and Melanie Jung

March, 2006

Abstract

A statistical model was developed to relate residential building shell leakage to building characteristics such as building height, floor area, floor leakage, duct leakage, and year built or the age of the house. Statistical regression techniques were used to determine which of the potential building characteristics best described the data. Seven preliminary regressions were performed to investigate the influence of each variable. The results of the eighth and last multivariable linear regression form the predictive model. The major factors that influence the tightness of a residential building are participation in an energy efficiency program (40% tighter than ordinary homes), having low-income occupants (145% leakier than ordinary) and the age of a house (1% increase in Normalized Leakage per year). This predictive model may be applied to data within the range of the data that was used to develop the model.

Table of Contents

List of Symbols

Symbol	Definition
NL	Normalized Leakage
Area	Floor area of the house in meters squared
H	Building height in meters
AT	Age of the house when it was tested
YB	Year when the house was built
FL	Existence of foundation leakage (0 or 1)
DL	Existence of duct leakage (0 or 1)
ε	Participation of the house in an energy-efficiency program (0 or 1)
β_x	Generic coefficient that in the regression
β_{Area}	The area coefficient
β_H	The height coefficient
β_{AT}	The age-tested coefficient
β_{YB}	The year built coefficient
β_{FL}	The floor leakage coefficient
β_{DL}	The duct leakage coefficient
β_ε	The e-program coefficient
$\beta_{x(std)}$	Generic standardized coefficient
I_{cz}	Vector of indicator variables for all the climates
I_{cold}	Indicator variable for the cold climate
$I_{mixed\text{-}humid}$	Indicator variable for the mixed humid climate
I_x	Generic indicator variable
β_{cz}	Vector coefficient corresponding to the climate vector
β_{cold}	Climate Coefficient for the cold climate
$\beta_{mixed\text{-}humid}$	Climate Coefficient for the mixed humid climate
I_{AT}	Indicator variable for age tested
I_{YB}	Indicator variable for year built
I_{FL}	Indicator variable for floor leakage
I_{DL}	Indicator variable for duct leakage
β_{IAT}	Coefficient of the age tested indicator variable
β_{IYB}	Coefficient of the year built indicator variable
β_{IFL}	Coefficient of the floor leakage indicator variable
β_{IDL}	Coefficient of the duct leakage indicator variable
$\beta_{adj.\ AT}$	Vector of coefficients corresponding to the climate vector, and adjusted for the age tested term
$\beta_{adj.}$	Vector of coefficients corresponding to the climate vector, and adjusted for multiple terms
ϕ_{Area}	Area factor in the predictive model for NL

ϕ_H	Height factor in the predictive model for NL
ϕ_{Age}	Age factor in the predictive model for NL
ϕ_ε	Energy efficiency program factor in the predictive model for NL (0 or 1)
ϕ_{Floor}	Floor leakage factor in the predictive model for NL (0 or 1)
ϕ_{LI}	Low-income occupant factor in the predictive model for NL
$\phi_{LI,Age}$	Low-income age adjustment factor in the predictive model for NL
$\phi_{LI,Area}$	Low-income area adjustment factor in the predictive model for NL
NL_{CZ}	A vector of constant terms, one for each climate, used in the predictive model for NL
size	Ratio of the floor area of a house to a reference area of 100 m^2
Age	Age of a house when it was tested (in years)
P^{Eff}	Percentage of houses in a dataset that participated in an energy efficiency program
P^{Floor}	Percentage of houses that have floor leakage (Floor leakage is defined as 1 if there is a possibility of leakage through the floor of the conditioned space as in a vented crawlspace or unconditioned basement, and 0 if there is no possibility of such leakage such as in a slab on grade house or a conditioned basement.
P^{LI}	Percentage of houses in a dataset that have low income residents

Introduction

The goal of this research was to create a model to relate residential building shell leakage to building characteristics such as building height, floor area, floor or duct leakage and the age of the house or the year it was built. A model was developed and statistical regression techniques were used to determine which of the potential building characteristics best described the data. The data used for the this project were from the residential leakage database compiled and maintained by the Energy Performance of Buildings Group at Lawrence Berkeley National Laboratory. Seven preliminary regressions were performed to investigate the influence of each variable. The results of the eighth and last multivariable linear regression form the predictive model.

Description of the database

The analysed database contains approximately 100,000 blower-door measurements[1] at single-family houses. The data were assembled from many different source organizations, therefore the building characteristics available for each house are not consistent throughout the database. The list of source organizations can be found in Appendix A. Most of the observations in the database contain the following core information: house floor area, test date, year built, participation in an energy efficiency program, and the shell leakage. A small number of houses have additional information such as the existence of a duct system, the type of floor or foundation construction, and the number of stories the house contains.

Quality of the data

The database does not contain equally distributed data that is representative of the U.S. housing stock because it was compiled from data that had already been collected in various research, certification and weatherization programs. By default the data contained in the database comes from houses that were chosen to be in one of these three types of programs. This means that we have a much higher percentage houses that participated in an energy efficiency program than there are in the housing stock at large. Our data are also not geographically uniform because each data source generally collected data from local houses so our data are somewhat clumped around our source sites. This paper will discuss the limiting characteristics of the database, and the model results will be applicable to the American housing stock within the limits defined by our data sources.

Non-Homogeneity of Data

The nature of the different sources leads to different information available for buildings in different parts of the country as well as geographical, construction quality, maintenance and operational differences between the buildings. The three largest contributors to the database are the Ohio Weatherization Program with more than 52,000 measurements, Alaska Housing Finance Corporation with almost 19,000 measurements, Energy Rated Homes with about 8,000 measurements and AKWarm (an energy-efficiency program in Alaska) with more than 5,000

[1] The leakage data from one of our sources, Energy Rated Homes with 8047 observations, were determined using both measurements of some houses and visual inspections of others. There was no indication in the dataset of which leakage values had been determined visually. We were assured by Energy Rated Homes that the fraction of visual inspections was small. In a more recent dataset from the same source the fraction of visual inspections was 4.6%. The visual inspection observations from the newer dataset were not included in the database. By applying the fraction of visual inspections from the new data set to the old data set, we approximate that the database contains 400 visual approximations of shell leakage.

measurements. More than 20 other organizations contributed the remaining measurements covering more than 30 states. We know that all the houses from the biggest contributor have low-income residents, since that was a requirement of participating in the Ohio Weatherization Program. That means the occupants of these houses earn below 125% of the poverty level. For the rest of the observations we don't have any information about the income of the occupants. Additionally, we have about 14,000 observations that we know were involved in an energy-efficiency program (e-program). That means some changes were made, either during the design phase or post-construction, to save energy. Some datasets did not offer information about energy efficiency programs so for these houses we don't know if they were involved in an e-program. For purposes of our regression we combined houses that were not in an e-program with those houses for which we didn't know their involvement with an e-program. There are also a significant number of observations that are missing other information such as year built, basement type, climate zone, etc.

Geographical Data Distribution

The data has a very significant regional bias since two thirds of the data are from Ohio and one quarter of it is from Alaska. Most of the Ohio data also come from low-income households since the major data source in Ohio is the Ohio Weatherization Program. In order to deal with this problem we separated the Ohio Weatherization Program from the rest of the data, and analyzed the two parts separately.

The bulk of the Alaska data was given to us after the analysis for this project was already finished. We ran the final regression (8) with and without the new data, and the results were not significantly changed, so we reported the results using all of the data. The preliminary regressions (1-7), however, do not include the new data.

Excluding the Ohio Weatherization Program data, the majority of the measurements are from houses located in Alaska, Rhode Island and Wisconsin. The abundance of data in Alaska is the main reason that the West Pacific Census Division is best represented. Data are also concentrated in Arizona, California and Washington, which are also in this Region. Rhode Island and Vermont together make up the second most sampled division, the New England Census Division. This is followed by the West Mountain Census Division, which consists of data from New Mexico, Arizona and Nevada. The South East Central Division is the worst represented division with fewer than 50 observations.

The U.S. Census Bureau collects U.S. housing stock data in a survey called the American Housing Survey (AHS). To create current statistics the Department of Housing and Urban Development interviewed 58,400 house owners in 1999. Chan et al. [2003] compared the database with these results. They pointed out that the floor areas of houses in the leakage database are generally smaller than those reported in the AHS. Also, the houses in the air leakage database are slightly older than houses in the AHS dataset.

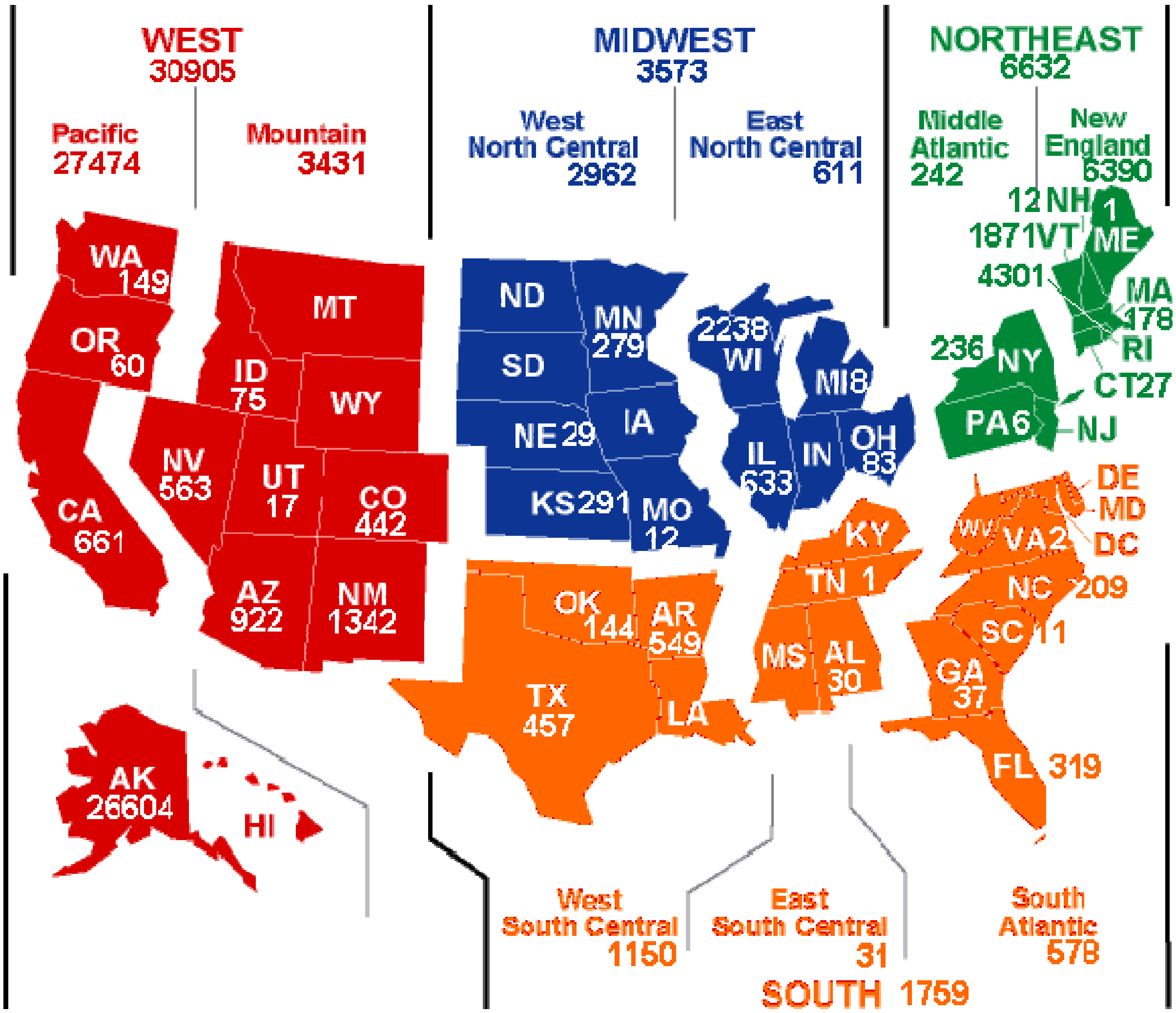

Figure 1: Geographic Distribution of Leakage Measurements in Database (2006) excluding the data from the Ohio Weatherization Program

Regression

Regression analysis is a statistical method where the mean of one or more random variables is predicted, based on other (measured) random variables. The leakage of a building is generally measured with a fan, and the data that is collected is the flow through the fan at a specific pressure, generally 50 Pascals. With this raw data it is difficult to compare differently sized houses with each other because there is usually more flow through a large house than a small one. Often, the raw data are normalized by converting the flow to an Equivalent Leakage Area [Sherman, 1995], and then normalizing the leakage area by floor area and height according to Equation 1, for Normalized Leakage as defined by ASHRAE [1988, 2005].

$$NL = 1000\left(\frac{ELA}{Area}\right)\left(\frac{H}{2.5m}\right)^{0.3}$$

Equation 1

When we look at the distribution of the Normalized Leakage in our database we see that it is not normally distributed, but that the distribution is closer to log-normal. This is expected because the Normalized Leakage is always a positive number. Regression analysis assumes that the data will be normally distributed so instead of regressing the normalized leakage, we regress the natural log of the normalized leakage. Figure 2 shows the distribution of our data on a log scale and a curve of a normal distribution with the same mean and standard distribution as our data. We can see that the distribution of NL is slightly skewed toward higher leakage values. After the regression we will transpose the equation and fitted parameters from log-space back into normal-space.

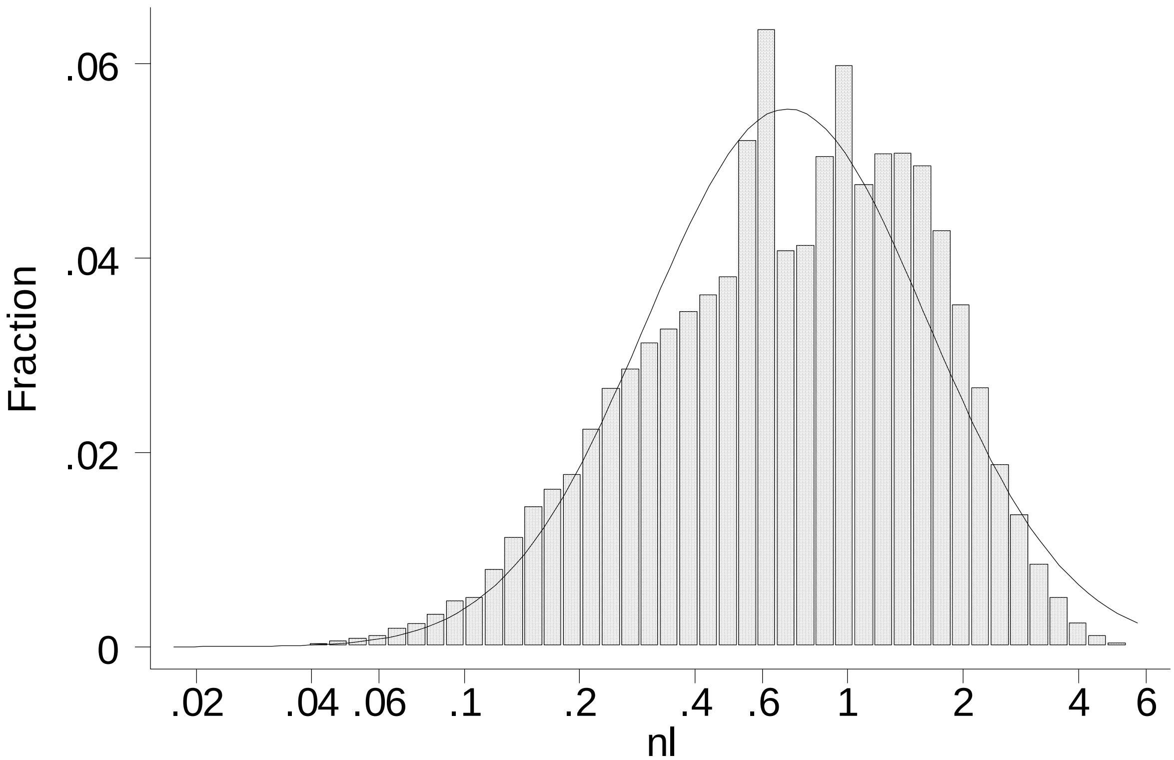

Figure 2: Distribution of the Normalized Leakage on a log scale

In analyzing the results of the regressions we look not only at the R squared of the regression as a whole, but also at the confidence interval which shows the ranges in which 95% of the values lie. Adding independent variables to a linear regression model will increase the value of R-squared for the regression unless the added variable is multicollinear with the existing variables. The effect can be accounted for by using the adjusted coefficient of determination (adjusted R-squared), which is always smaller than R-squared and can also be negative. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance, and is used in this analysis whenever R-squared is referred to.

The individual coefficients can be examined by using the t-value and P-value. The t-test (yielding the t-value) is a statistical test of whether the slope of a regression line differs significantly from 0. In statistics, a result is significant if it is unlikely to have occurred by chance, where, in reality, the independent variable being examined has no effect. Thus, the larger the t-value for a particular coefficient, the larger the statistical difference between that slope and zero.

The P-value is the significance level of the t-test or the maximum probability of accidentally rejecting a true null hypothesis. (The null hypothesis in this case is the hypothesis that the particular coefficient is equal to zero.) The smaller the P-value, the more significant the result is said to be.

In the analysis we compare the significance of the variables, one to another. In order to do this it is first necessary to standardize the variables by subtracting the mean from each value and then dividing by the standard deviation of the distribution. In this way the units of each variable are removed, and the distribution of each variable is centred on zero, with a standard deviation of one.

Data analysis and processing

Data processing was initiated by verifying the plausibility of each data point. This means unrealistic data values (such as building year earlier than 1600) were deleted from the database. Afterwards the available information was investigated to see which independent variables are qualified for the regressions. Each of the variables is described in the later part of this section.

Error Correction

The following acceptable data ranges were applied to the data:

- floor area — from 30 to 1,000 square meters
- building height — taller than 1.79 meters
- year built — 1,600 or newer
- cfm50 — from 100 to 20,000
- year tested — 1980 or more recent,

The minimum year tested date was set at 1980 because both big companies which manufacture blower door testing equipment in the United States were founded in the early 1980's.

We identified and fixed a mistake that had been made in the data entry of the shell leakage of approximately 8,000 data points that had been noticed in previous investigations of this database.

Another data challenge was to assign climate zones (since we thought leakage might vary with climate). Many of our data had spelling errors in the location information, or inconsistencies between the city, state and zip code. These errors were fixed where possible and then a climate zone was assigned to each record based on the Building Science Corporation's climate map, see Figure 3.

Figure 3: Climate Zones defined by Building Science Corporation

Uneven Data Distribution

Houses in the leakage database do not statistically represent the characteristics of the housing in the US as a whole because of two main reasons.

1. Data were contributed voluntarily by home weatherization contractors and research organizations from around the country, and some contractors contributed much more data than others.

2. Most of the data were gathered as part of programs to target particular classes of homes, for example, "low-income" homes that were tested as part of a weatherization program, and "energy-efficient" homes that were tested to check compliance with air infiltration targets of the energy programs.

The Normalized Leakage distribution is shown in Figure 4 for the four best represented climate zones, excluding houses with low-income residents and e-program houses. The tail of the cold data with NL greater than 2.5 was not visible on the graph when all data was shown together. The tail consists of 13 data points with NL greater than 2.5 and less than 4.5. The very cold climate also has a tail of 4 data points with NL greater than 2.5 and less than 3.0. Data from a particular data source[2] is visible in the cold and very cold climates where we see a few discrete values of NL for a portion of the data. In this particular data set we were not able to obtain raw data, but only data that had already been categorized into leakage classes. The cold climate, with median NL of 0.59, has leakier houses than the other three climates with median NL values of 0.46, 0.22, and 0.40 for the very cold, sub arctic and mixed & hot dry climates respectively. Quartile values are summarized in Table 1.

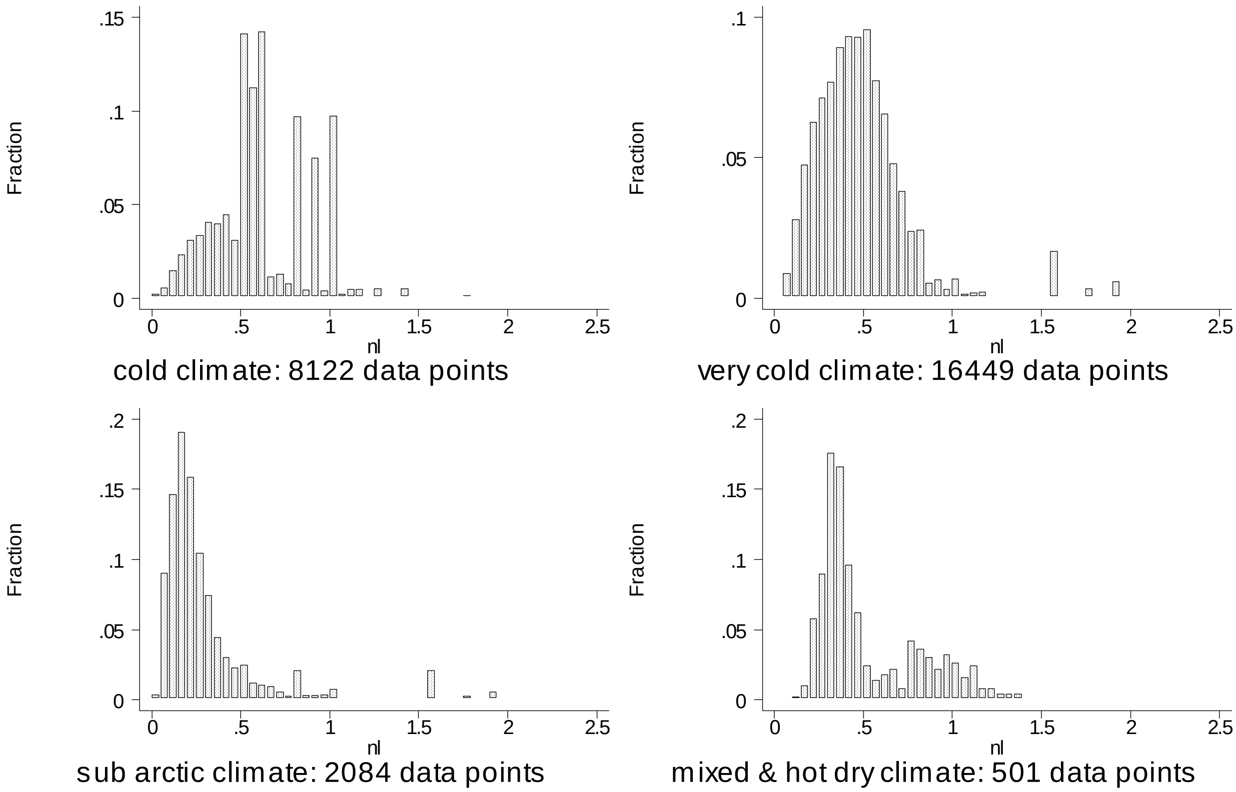

Figure 4: Normalized Leakage distribution for the best represented climates, excluding houses with low-income residents and e-program houses

Climate	**25th Percentile**	**50th Percentile**	**75th Percentile**
Cold	0.48	0.59	0.83
Very Cold	0.32	0.46	0.61
Sub Arctic	0.15	0.22	0.34
Mixed & Hot Dry	0.33	0.40	0.75

Table 1: Normalized Leakage quartiles for the best represented climates, excluding houses with low-income residents and e-program houses

[2] Energy Rated Homes containing 8047 observations, all tested prior to 1994.

More than half of the data comes from a low-income weatherization program in Ohio, making this type of house over-represented in our dataset. Figure 5 shows the distribution of normalized leakage values in ordinary, e-program and low-income houses. The tail of the ordinary data has been graphed separately because it was not visible when all the data was shown together. The e-program data also has a tail of 2 data points with NL greater than 2.5 and less than 2.7.

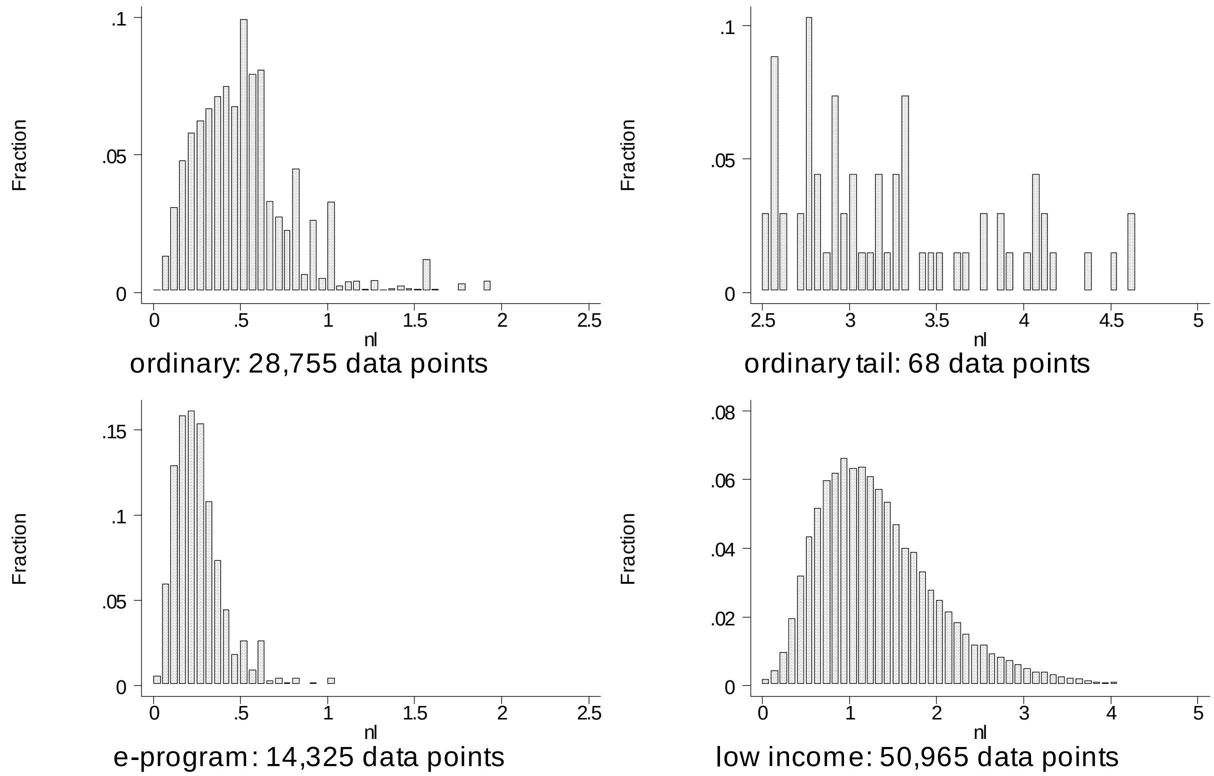

Figure 5: Normalized Leakage distribution for all ordinary houses, e-program houses, and houses with low-income residents

We can see from the graph, and also from the quartile values in Table 2 that e-program houses with a median NL of 0.25 are tighter than ordinary houses with median NL of 0.50 and low income houses with a median NL of 1.24. We have a nice distribution of low income houses because of the huge volume of data.

House Type	25th Percentile	50th Percentile	75th Percentile
E-Program	0.17	0.25	0.34
Ordinary	0.33	0.50	0.65
Low-Income	0.85	1.24	1.74

Table 2: Normalized Leakage quartiles for three types of housing: e-program houses, houses with low-income residents and ordinary houses (those which are not part of either of the previous groups)

The ordinary data are much less uniform due to the discrete values from the Energy Rated Homes source. We have fewer e-program houses than ordinary houses in our database, therefore we believe that the additional tightness of the e-program houses can be captured in an e-program variable. We will try the same approach with the low-income data in regression 5, but do not anticipate success since there is so much data from the Ohio Weatherization Program. As part of the exploratory analysis we divide the houses into two broad classes, “low income” and “ordinary”, and analyze the two classes separately. We will examine each of the regression variables to see if it has a different trend in the low-income versus ordinary data. However, since all of our “low income” houses are in Ohio we have to assume that “low income” houses in Ohio differ from the Ohio housing stock in the same way that “low income” houses in other states differ from the housing stock in those states. This may or may not be a good assumption.

Variables Investigated

Eight variables were investigated: foundation type, year built, house age when tested, low-income residents, participation in energy efficiency program, floor area, climate zone and the existence of a duct system. Six of these variables were included in the final model. Each of the variables is described below.

Foundation Type

Sherman and Dickerhoff [1998] point out that the Normalized Leakage of houses with a slab-on-grade foundation is significantly less than for houses with a crawlspace or an unconditioned basement. Unfortunately, we know the under floor construction for fewer than 10% of the houses in the database. The 2003 AHS reports the presence of a slab, basement, or crawlspace to be about 29%, 43%, and 27% in the US housing stock. However, the survey did not differentiate between conditioned and unconditioned basements. The leakage database, in comparison, reports about 9%, 53%, and 34% of the houses having a slab, basement, or crawlspace, respectively.

House year-built and testing age

New homes tend to be much tighter than old homes because of improved materials (e.g. weather-stripped windows), better building and design techniques, and lack of age-induced deterioration (e.g. settling of foundation). This trend has been reported by Sherman and Dickerhoff [1998] who observed substantial reduction in leakage in homes built after 1980. We attempted to determine separately the reduction in leakage due to improvements in building technologies, and the aging effect due to deterioration by using a “year built” variable to capture the change in construction techniques over time, and an “age tested” variable to capture the effect of aging on building leakage. Because all the houses were tested within a span of 20 years (1981 to 2004), these two variables are not completely independent. Houses that are less than 10 years old are very likely to be built after 1980, and houses that are older than 20 years old are very likely to be built before 1980. Because of this it is very difficult to separate the effect of materials or construction technique improvement which experienced a sharp reduction in leakage around 1980 and a deterioration effect due to aging. Most of the observations in the database include information about testing year and year built but not all. The age tested data are calculated by subtracting year built from testing year. These two variables were found to be correlated so only one could be used in the final model. Testing age was used in the model and year built was discarded.

Low-Income

Chan et al. [2003] point out that low-income houses have much higher leakage areas than ordinary houses, regardless of year built and floor area, and we also see this in Figure 5. We examine this in the analysis. As previously mentioned, all of our low-income data comes from a source called Ohio Weatherization Program, and all of this data are located in Ohio.

Energy-Efficiency Programs

Houses that are participants in energy-efficiency programs are designed to be especially air tight to save thermal conditioning costs. Nearly fifty percent of the non low-income data are from energy-efficiency programs in 31 different states. All of the data from New Mexico, Kansas and Pennsylvania are from energy-efficiency programs. The fraction of houses in an energy-efficiency program in the database is much higher than observed nationally. This is because blower-door measurements are often used for the energy analysis that is commonly performed on houses participating in energy-efficiency programs.

Floor Area

Normalized Leakage is normalized by floor area, so we don't expect a relationship between Normalized Leakage and floor area, but Chan et al. [2003] identified that Normalized Leakage is a function of floor area among houses that were built before 1995 so we investigate this variable. A relationship between NL and floor area would suggest that ELA should be normalized not by 1/Area, but by something more complicated.

Climate zones

We expect that the climate has an important influence on the leakage area of residential buildings. Houses in harsh climates should be tighter because increased infiltration due to stack effect will result in more discomfort since the infiltrating air is cold in winter. In milder climates we expect more leakage because there is a lower monetary and comfort incentive to build tight houses.

The International Energy Conservation Code (IECC) defines 17 Climate Zones [ICC, 2003]. When our data are classified into these climate zones some zones contain very few or no data points. We need to group some climate zones together in order to do a meaningful climate analysis. Building Science Corporation uses a simplified set of 7 climate zones which can be directly mapped to the 17 IECC climate zones [BSC 2005]. See Figure 3 for a map of the climate zones and Table 3 for a listing of the number of observations that our dataset has in each zone. Although the marine climate still has a small number of data points, the other climates seem to have sufficient data so we begin our climate analysis with these climate zones, and we will later combine them if necessary.

Climate	**Total Number of Observations**	**Observations from Ohio Weatherization Program**
sub arctic	3,736	
very cold	23,202	
cold	55,154	44,956
dry	3,362	
mixed-humid	7,285	6,009
hot-humid	810	
marine	293	
unknown	276	

Table 3: Number of Observations in each Climate Zone

Duct system

The presence of a thermal distribution system can add significant leakage. Sherman and Dickerhoff [1998] report that leaks measured separately from duct systems account for almost 30% of the total leakage of the house. The American Housing Survey classifies heating equipment into several types, but the two that use ducts as part of the system are warm-air furnaces and electric heat pumps. They represent 60% and 10% of the total housing stock respectively, meaning that 70% of the housing stock contains a duct system. In the air leakage database, there are approximately 5000 data points that record the presence of absence of duct systems, with 74% of those reporting the presence of duct systems. In the end this variable was not used in the model.

Result of data analysis and processing

Based on these exploratory analyses, we conclude that house type (low-income, ordinary or energy-efficient), year built, age, climate, foundation construction and floor area all potentially influence the leakage of a house. As these factors are not completely independent of one another, more detailed analysis is required to determine how each one is associated with leakage. This more detailed analysis will be done in the next section by separating the data into categories (low-income and ordinary data) and carrying out the regression analysis.

Regression Analysis

Eight sets of regressions were run in this analysis. Each set consisted of one regression of the low-income dataset and another regression of the ordinary house data set until the two datasets are combined for the later regressions. We started with a basic model and made changes in each successive regression until we arrived at the final regression, which we think best describes the data. Regression 8 was performed with and without the new data, and is described in detail in the Predictive Model section.

The first regression used only those variables for which each observation had a value. The low-income data and ordinary data are analyzed separately for all regressions until Regression 5. In the second regression we develop a method for including observations with missing data. Regression 3 adds the duct leakage and floor leakage variables. Regression 4 adds the energy efficiency program variable. Regression 5 combines the low income and ordinary house data into one model. In Regression 6 several variables are removed in turn to see if they are necessary to include in the model. Regression 7 examines the climate zones in detail.

As described in the Regression section of the Introduction, our data shows that the Normalized Leakage has logarithmic distribution. Therefore, instead of creating a linear model for NL, we will develop a linear model for the natural logarithm of NL, which then becomes an exponential model for NL as we will see in the Predictive Model section.

Results of the regressions using standardized variables can be found in Appendix B. The results of non-standardized regressions are summarized in Appendix C, and the full information on each regression (including the t and P-values) can be found in Appendix D. In general, the magnitude of the t-values are large (the sign always follows the sign of the coefficient) and the P-values are zero. Where this is not the case it is noted in the text.

Regression 1

We start with a basic model which uses only observations where all the parameters are known: floor area, building height, age the house was when tested and the year it was built. We regress the low-income house data and the ordinary house data separately.

Low Income Data

The Ohio Weatherization Program data come exclusively from low-income households. There are more than 50,000 observations in this dataset that contain values for each of the necessary parameters. All the observations are in only two climate zones. The indicator variables for each of the climate zones is set to 1 if the observation is in that particular climate zone and zero if it is not. There were no observations with unknown climate in this dataset. The model for these data is then the following:

$$\ln(NL) = (\beta_{cold} \cdot I_{cold} + \beta_{mixed\text{-}humid} \cdot I_{mixed\text{-}humid}) + \beta_{Area} \cdot Area + \beta_{H} \cdot H + \beta_{AT} \cdot AT + \beta_{YB} \cdot YB$$

Equation 2

As previously stated, in order to compare the variables, one to another, it is first necessary to standardize the variables by subtracting the mean from each value and then dividing by the standard deviation of the distribution. In this way the units of each variable are removed, and the distribution of each variable is centred on zero, with a standard deviation of one.

Aside from the two constants for the climate, the standardized age coefficient, 0.98, has the largest magnitude of the coefficients in this first regression showing that age has the most influence on the leakage of a building. The algebraic sign is positive indicating that the older the building is the leakier it is like expected to be.

The second biggest coefficient is the year the house was built with a value of 0.8. The positive algebraic sign means a positive relationship between the year a house was built and its leakage area. The later the building was built the leakier it is. This is in contrast to what we expect. The coefficient of house age when tested indicates that older houses are leakier, which is what we expect. This leads us to look at these parameters in more detail.

The correlation matrix shows a correlation of 94% between the influences that each of these variables have on the Normalized Leakage. Since all houses in our database were tested more recently than 1980 there is a strong correlation between date the house was built and the age of the house when it was tested. Because of this, only one of these variables should be used in the final regression. We will investigate which is the better variable to use in regression 6.

We expect that the floor area will have a relatively small influence on the Normalized Leakage since the leakage is normalized by floor area, as well as by building height. The standardized floor area coefficient is –0.24, which is the second smallest coefficient, confirming our expectations. The non-standardized floor area coefficient is -0.0044 [1/m2], which means that the logarithm of NL actually decreases by 0.0044 for each additional square meter of building floor area.

The coefficient for the building height (-0.002) is the smallest in regression 1, and has a standard deviation larger than the value of the variable (0.003). The t-value is small, and the P-value is 0.572, indicating low statistical significance for this variable in this regression.

An examination of the building height data reveals that it has only two discrete values (2.5m and 3.75m) in this part of the dataset because no data on building height or number of stories were given from the Ohio Weatherization Program directly. The floor area was assigned according to the following assumptions: Single-family buildings with a floor area smaller than 92 m^2 were assigned the height of a a 1-story building (2.5 meters). Buildings with floor area 92 m^2 and bigger were assigned the height of 1.5-stories (3.75 meters). These assumptions are permissible because the Normalized Leakage is only weakly dependent on building height due to the height term's exponent of 0.3. The area for the low-income data set varies between 46 m^2 and 1,030 m^2. This results in a distribution of 40% 1-story buildings and 60% 1.5-story buildings. Because this coefficient has a standard error that is bigger than the coefficient itself, the building height will not be included as one of the variables in the next regression.

We calculated the constants separately for the two climate zones containing data in Ohio, cold and mixed-humid. The P- and t-value of both coefficients point out significance for both constants. As expected the constant term for the cold climate is smaller. That means the Normalized Leakage is smaller in the colder climate indicating tighter houses. Each of the individual climate zones (cold and mixed humid) has a higher adjusted R-squared and smaller root mean square error (MSE) than the two climate zones together.

The R-squared for this regression indicates that this model only describes 21.5% of the data variability.

Ordinary Data

For the houses that are not part of the Ohio Weatherization Program, the model becomes slightly more complicated because we have a wider range of climates so we replace the two constant terms with a sum of eight constant terms:

$$\ln(NL) = \Sigma(\beta_{cz} \cdot I_{cz}) + \beta_{Area} \cdot Area + \beta_H \cdot H + \beta_{AT} \cdot AT + \beta_{YB} \cdot YB$$

Equation 3

$I_{climate}$ is a set of climate indicator variables, where, I_x is equal to 1 for the x^{th} climate and 0 for all the others. There are seven climates in this dataset, plus an additional climate variable for those where the climate is unknown.

All of the coefficients change when we use this new dataset. The height coefficient (0.14) is two orders of magnitude bigger than it is in the low-income-data and is the second biggest coefficient in this regression. The standard error is less than 2% of the coefficient. This is probably because we have real data for this variable instead of discrete data correlated with floor area. The algebraic sign is positive that means the higher the building the bigger the Normalized Leakage area.

The coefficients of all the other variables, floor area, "age tested", and "year built", decrease using the new dataset showing that the dependence of the Normalized Leakage on these variables decreases in the ordinary data.

The climate coefficients are the same order of magnitude as in the regression of the low-income data (low-income data: I_{cold} = -57.22 and $I_{mixed\ humid}$ = -57.05; ordinary data: I_{cold} = -20.4 and $I_{mixed\ humid}$ =-20.1).

The adjusted R-squared is in this regression shows that nearly 36% of the variability in the data is described by the model. It is almost twice that of the low-income data but it still describes less than a half of the variability. Although separate regressions for each climate zone were investigated, a separate equation for each climate zone is not an option because of the vastly different number of observations across the climate zones. Moreover we want to have one model, which is as simple as possible, so we continue with the combined model, but modify it in the next regression to include observations with missing information for some of the variables. (In regression 1 these observations were dropped from the regression.)

Regression 2

In regression 2 we remove building height when we regress the low-income data and we include observations with missing values by adding an indicator variable for each of the problem variables. The indicator variables (I_x) are 1 if the data are known and 0 if the data are unknown, just as for the climate variables. The following term is substituted for each term in the previous equation that contained missing data:

$$\ldots + \beta_x \cdot X \cdot I_x + \beta_{Ix} \cdot (1 - I_x) + \ldots$$

Equation 4

Low Income Data

In the low income data we use indicator variables for the year built and age tested terms. The building height variable is dropped, and the floor area variable contains no missing data. This regression uses about 250 more observations than were used in regression 1. Because the age tested variable is calculated by subtracting year built from year tested, and in this dataset the year tested variable was complete, all observations where year built is missing are also missing the age tested. Therefore, we need only one indicator variable for both variables ($I_{AT}=I_{YB}$).

$$\ln(NL) = (\beta_{cold} + \beta_{mixed\text{-}humid}) + \beta_{Area} \cdot Area + \beta_{AT} \cdot AT \cdot \mathbf{I_{AT}} + \boldsymbol{\beta_{IAT}} \cdot \mathbf{(1\text{-}I_{AT})} + \beta_{YB} \cdot YB \cdot \mathbf{I_{YB}} + \boldsymbol{\beta_{IYB}} \cdot \mathbf{(1\text{-}I_{YB})}$$

Equation 5

The fit is nearly the same for this regression as it was for regression 1 when we removed the building height variable from the regression. This means the assumptions about the indicator variables that we made were good.

Ordinary Data

Many observations in the ordinary data set have missing values for year the house was built and testing year. We can include these observations in the regression by using indicator variables as shown in Equation 6. Here the indicator variables I_{AT} and I_{YB} are not equal to each other.

$$\ln(NL) = \Sigma(\beta_{cz} \cdot I_{cz}) + \beta_{Area} \cdot Area + \beta_H \cdot H + \beta_{AT} \cdot AT \cdot \mathbf{I_{AT}} + \boldsymbol{\beta_{IAT}} \cdot \mathbf{(1\text{-}I_{AT})} + \beta_{YB} \cdot YB \cdot \mathbf{I_{YB}} + \boldsymbol{\beta_{IYB}} \cdot \mathbf{(1\text{-}I_{YB})}$$

Equation 6

The ratio of the parameters coupled to each variable can be used to estimate the average value of the missing data. If we assume that there is no inherent difference in the missing and non-missing data with respect to the variable, i.e. $\beta_{x(missing)}=\beta_x$, then β_{Ix} is equal to the product of β_x

and the average of the missing values, and we can calculate the average of the missing variables as follows:

$$\frac{\beta_{Ix}}{\beta_x} = \bar{x}_{missing}$$

Equation 7

The average of the missing values, calculated in this way, is 128 years for the age of the house when it was tested and 1987 for the year built. These numbers are different from the average of the known ones ($\overline{AT} = 10$ and $\overline{YB} = 1993$). The difference between the known and missing values for both of these variables intuitively makes sense because an occupant is less likely to know the year a house was built the older it is. What is surprising is that there are 118 years of difference between the values of age tested, and only 6 year of difference between the values of year built. The averages of the low-income-data show a similar behavior. The calculated averages for testing age and year built are 116 years and 1857. The averages of the known values are 52 for age tested and 1942 for year built. In this dataset the two variables have closer to the same number of years difference between known values and missing values (64 years for the age tested variable, and 85 years for the year built variable). These results do not have any relationship to how the age of the house affects the Normalized Leakage, and it is possible that the results are skewed by the correlation between the age variables, but the result is interesting nonetheless.

The number of observations is 50% larger in this dataset by using all observations. Also the adjusted R-squared increased by nearly 5%.

Regression 3

In this step we added the variables and indicator variables for duct and floor leakage. The data from the Ohio Weatherization Program contained no information about existing ducts or basement type. So we only look at the ordinary data.

Including these new variables, the model becomes:

$$\ln(NL) = \Sigma(\beta_{cz} \cdot I_{cz}) + \beta_{Area} \cdot Area + \beta_H \cdot H + \beta_{AT} \cdot AT \cdot I_{AT} + \beta_{IAT} \cdot (1-I_{AT}) + \beta_{YB} \cdot YB \cdot I_{YB} + \beta_{IYB} \cdot (1-I_{YB})$$
$$\boldsymbol{+ \beta_{FL} \cdot FL \cdot I_{FL} + \beta_{IFL} \cdot (1-I_{FL}) + \beta_{DL} \cdot DL \cdot I_{DL} + \beta_{IDL} \cdot (1-I_{DL})}$$

Equation 8

Both variables are assigned a value of 1 if there is known leakage from the designated area (through the floor or through the duct system) and a value of 0 if there is no leakage through this area. There is no floor leakage in a slab on grade house or a house with a conditioned basement. Similarly, a house with no duct system cannot have duct leakage. Numbers between 0 and 1 may be assigned to these variables to denote a probability of duct or floor leakage when analyzing a large homogeneous dataset, or a percentage of full leakage on a case by case basis. In our dataset the variable FL is assigned the value 1 if the foundation type is crawlspace or unconditioned basement, 0 if the house is slab on grade or has a conditioned basement, and FL gets the value 0.5 if the foundation type is unknown or is a combination of foundation types. Similarly, the duct leakage variable has a value of 1 if the house has a duct system, 0 if it does not have a duct system.

Compared to Regression 2, the adjusted R-squared increases and the root mean square error goes down. This is expected since we have added two additional degrees of freedom to the model. The climate and year built variables change slightly. The three variables for the floor area, the building height and the indicator variable of the age are become closer to zero.

Only the coefficient for the age tested, β_{AT}, changes significantly. Again, this could be caused by the correlation of the age variables which will be investigated later.

The new floor leak variable has P- and t-values (t-value = -2.65 and P-value = 0.8%) that show it to be not well defined in this regression. This is probably because the presence or absence of floor leaks is known for just under 25% of the data.

The variable for the duct leakage is shown as significant by the P- and t-values. The negative algebraic sign of the coefficient (- 0.139) is surprising. It means that buildings without duct systems are leakier than buildings with duct systems. This is opposite to what is expected. Ducts often run through an opening in the building shell to an air handling unit that is located outside of the conditioned space. Therefore, the result that houses with ducts are tighter than those without is unbelievable. 85% of the buildings with duct information are from energy efficiency programs. The e-program variable was not considered in this regression so it is likely that building tightness due to e-program shell improvements is ascribed to the duct variable in this regression.

Regression 4

In this regression we introduce another variable which will capture the effect of participation in an energy efficiency program on the Normalized Leakage. The e-program parameter has a slightly different form because in the data we cannot tell the difference between a house that is truly ordinary from one in which the participation in an energy efficiency program is unknown. Therefore we only have one set of input values not two.

$$\ln(NL) = \Sigma(\beta_{climate} \cdot I_{climate}) + \beta_{area} \cdot area + \beta_H \cdot H + \beta_{AT}\, AT \cdot I_{AT} + \beta_{IAT} \cdot (1-I_{AT}) + \beta_{YB} \cdot YB \cdot I_{YB} + \beta_{IYB} \cdot (1-I_{YB}) + \beta_{FL} \cdot FL \cdot I_{FL} + \beta_{IFL} \cdot (1-I_{FL}) + \beta_{DL} \cdot DL \cdot I_{DL} + \beta_{IDL} \cdot (1-I_{DL}) + \boldsymbol{\beta_\varepsilon} \cdot \boldsymbol{\varepsilon}$$

Equation 9

In this model there are 19 independent variables and 19 fitted parameters (β_x). Eight parameters for the climate zones; one each for area, building height and energy programs and two each for age tested, year built, foundation type and ducts.

Compared to Regression 3 the adjusted R-squared goes up and the root mean square error goes down, as they must because an additional variable was added. The coefficient for the e-program variable is negative (-0.22) so the Normalized Leakage for houses participating energy efficiency programs is smaller, all other variables being equal. It is not surprising that energy program houses are tighter since they are built to save energy.

The duct leak coefficient is still negative, but a slightly higher t-value shows a bit more significance in this regression than in regression 3. The natural logarithm of NL for buildings with duct system ($\overline{\ln(NL)}_{DL=1} = -1.36$) is smaller than the logarithm of NL for buildings without ducts ($\overline{\ln(NL)}_{DL=0} = -0.94$). Apparently the tightness of these particular houses that have a duct system is not sufficiently accounted for in the e-program variable, and the regression assigns the additional tightness to duct leakage, although we know this to be impossible. It follows that the variable is not qualified to be used in our mode because our data are skewed with respect to this variable.

The variables for the floor area, the year built (including indicator variables), the duct variable, and climate variables don't change much from the previous regressions. They all become slightly closer to zero, as we expect when we add more degrees of freedom to the model.

In this regression the coefficient for the variable of the house age when tested is negative. This is probably due to the 92% correlation between the age tested and the year built variables. The P-value of the age tested coefficient is 0.3% in this regression compared to 0 in the last regression which indicates that it was better defined in the previous regression.

The floor leakage variable in this regression has an even larger P- value (17%) and a smaller t-value (-1.37) than in regression 3 indicating even lower significance. We drop this variable from the model, but will revisit it again in Regression 8 when we use a different method for including variables with many missing data points.

Regression 5

In this regression we develop a way of combining the low-income data and the ordinary data into one model. The fitted parameters relating to floor area, building height, year built and age tested are dissimilar between the ordinary dataset and the Ohio Weatherization Program dataset. We combine the two datasets into one regression in the simplest way possible, by adding a single parameter (LI) to the model of regression 4 as shown in Equation 10.

$$\ln(NL) = \Sigma(\beta_{cz} \cdot I_{cz}) + \beta_{Area} \cdot Area + \beta_H \cdot H + \beta_{AT} \cdot AT \cdot I_{AT} + \beta_{IAT} \cdot (1-I_{AT}) + \beta_{YB} \cdot YB \cdot I_{YB} + \beta_{IYB} \cdot (1-I_{YB}) + \beta_{FL} \cdot FL \cdot I_{FL} + \beta_{IFL} \cdot (1-I_{FL}) + \beta_{DL} \cdot DL \cdot I_{DL} + \beta_{IDL} \cdot (1-I_{DL}) + \beta_\varepsilon \cdot \varepsilon + \boldsymbol{\beta_{LI} \cdot LI}$$

Equation 10

The LI variable is 1 or 0 depending on whether or not the data comes from the Ohio Weatherization Program dataset. Like the energy program issue, we only have one variable. It is known that those in the Ohio Weatherization Program are low-income houses, but the ordinary dataset may also contain some low-income houses.

In this regression the adjusted R-squared increases to 64% from 51% in regression 4. It seems that this is an acceptable way to combine the two datasets. After closer examination we see that all of the climate variables and the year built variables have P-values greater than zero and t-values smaller than 1. We decide to continue analyzing the data from the two datasets separately and find another way to combine them for the predictive model at the end.

Regression 6

In previous regressions we saw that the age tested variable and the year built variable are correlated, so only one of them is appropriate to use in the model. In this regression we remove the year built variable and age tested variable in turn to see which of these variables better describes the data. We remove the floor leakage variable in this regression because regression 4 indicated that it was not significant.

The adjusted R-squares of the two regressions (.49 using age tested and .50 using year built) are very similar, and are very close to the adjusted R-squared of regression 4 (.51) so the model works just as well without the variable for the floor leaks and without one of the age variables. The standardized coefficients in these two regressions are very similar to each other, and are not much different from the coefficients in regression 4, except for the age tested coefficient and the climate coefficient in the regression that uses age tested variable. In this regression the leakage is explained by the age tested, rather than the climate. The correlation matrix for this regression does not show a correlation between age tested and climate, whereas the correlation matrix of the next regression shows a high correlation between the year built variable and the climate coefficients. It makes sense that the climates are related to the year the house was built because different areas of the country experienced building booms at different times. For this reason we choose to use age tested as the variable that we will continue to use in the following regressions. It is important to remember, however, that this coefficient really describes both an aging effect, and an improved design and materials effect so the predictive model should not be used to predict into the future, but only to back-forecast the leakage of existing houses.

Regression 7

Now the developed model fits the data fairly well. There is no variable with a t-value smaller than twenty. All the P-values are zero. Now we turn our attention to the climate zones. Some of the climates have very few observations compared to the other climates. The coefficient values are shown in Table 4 with the tightest houses at the top of the table, and the leakiest at the end.

Climate Zone	Number of Observations	Coefficient Value
sub arctic climate	3736	-1.39 ± .02
marine climate	293	-1.14 ± .03
mixed and hot dry climate	3362	-1.05 ± .02
cold climate	10,198	-0.99 ± .02
very cold climate	23198	-0.88 ± .02
mixed humid	1276	-0.66 ± .02
unknown climate	276	-0.54 ± .03
hot humid climate	811	-0.46 ± .03

Table 4: Number of observations and coefficient value per climate zone from regression 6 using the age tested variable.

In regression 7 we combine the climate zones to eliminate some of the climate zones with low numbers of observations, and to align the climate zone boundaries with areas that have similar building practices. This is particularly important for Alaska, which we know to have different building practices, for instance, the garage is conditioned. We create a new climate called Alaska containing all the observations from this state. These observations are a subset of the former sub arctic and very cold climates.

The second new climate is the cold climate. Because of the large number of observations the cold climate is representative enough to stay alone. The few observations from the sub arctic (279 observations) and very cold (52 observations) climate that are not situated in Alaska fit in this climate too.

The new humid climate includes the former hot and mixed humid climates. This climate is located in the south east of the country.

The remaining climates are summarized in the dry climate. The dry climate covers the west and south west of the country. Although the marine climate is not particularly dry, the building practices in this climate are similar to building practices all over the west of the country.

We expect the houses in more severe climates to be tighter than those in mild climates, and this is the case with the tightest houses in Alaska as shown in Table 5. We found that the climates that the model predicts to be tighter are those climates where a higher proportion of our data comes from e-program houses. We will look at the climate coefficients in more detail when we do the final regression, because it may change.

Climate Zone	Number of Observations	Coefficient Value
Alaska climate	26,603	-1.31 ± .02
dry climate	3,655	-1.10 ± .02
cold climate	10,529	-1.03 ± .02
humid climate	2,087	-0.62 ± .02
unknown climate	276	-0.59 ± .04

Table 5: Number of observations and coefficient value per climate zone from regression 7

Results of the preliminary regression analysis

Regression 7 describes 45% of the data by using five constant terms (climate coefficients), five variables and two indicator variables. The logarithm of Normalized Leakage decreases by 0.0014 per added square meter. If the building height increases by one meter the ln(NL) increases by 0.08. This value is plausible since the surface of the building shell increases with increasing height.

The leakage expressed in ln(NL) increases by 0.011 per year. This effect contains both aging and improvements in the quality of construction. Since the age and year built data are correlated it is not possible to separate these effects.

Our model shows that houses with ducts are tighter than those without ducts. This is counter-intuitive and physically unrealistic so we look for another explanation. 85% of the houses in our database that have ducts are also e-program houses. Perhaps these e-program houses with ducts are slightly tighter than the average e-program house, and that tightness is erroneously explained by the ducts variable although it has nothing to do with the ducts. Therefore, the duct variable will not be used to form the predictive model.

Predictive Model

The intended purpose of this model is to predict the Normalized Leakage of a set of houses in the U. S. If we had statistically representative and complete data we would simply fit our model to the data and get the parameters of interest. Unfortunately we have missing data and we have unrepresentative data. We will therefore try to stratify the analysis procedure to maximize the value of the data: We will develop a core model and then regress the residuals of the model to fit the other parameters. Indicator variables are no longer necessary with this new strategy.

Development of the Predictive Model

Since we have so much data in Alaska, we ran the analysis on the Alaska data and the continental US separately to see if there was a difference between the data that we have for these regions. (Results of these regressions are shown in Table 6.) There was not a significant difference so we decided that no stratification was required for Alaska in the regression 8 analysis.

Regression 8: The Core

For all the data we are using we have information on the leakage, the climate, the area and height of the house and whether it is in an e-program. We therefore will use all of the *non-low income* data and fit it to what we call our core model:

$$\ln(NL) = \overleftrightarrow{\beta_{cz}} + \beta_{Area} \cdot Area + \beta_H \cdot H + \beta_\varepsilon \cdot \varepsilon \qquad \textbf{Equation 11}$$

We have excluded the low-income data from the core because the low-income data we have comes only from one state and therefore may be biased for a variety of reasons. We have excluded age from the core because only about half the non-low income data has appropriate age information *and* the leakage of houses where the age is known is different from the leakage of houses where the age is not known.

Regression 8a: Adjustment for age

To estimate the age parameters we will first consider the non-low income data for which we have age information. We will regress the residuals of regression 8 against the age variable. Specifically,

$$\ln(NL) - (\overleftrightarrow{\beta_{cz}} + \beta_{Area} \cdot Area + \beta_H \cdot H + \beta_\varepsilon \cdot \varepsilon) = const + \beta_{AT} \cdot AT$$

Equation 12

The age coefficient is, in fact, the one we wish to use in our predictive model, but in order to keep the mean leakage unchanged between the core and this expression it is necessary to subtract off the impact that this new parameter has from the climate term which is the parameter times the average age of the homes in the dataset: $\overleftrightarrow{\beta_{adj.AT}} = \overleftrightarrow{\beta_{cz}} - \beta_{AT} \cdot \overline{AT}$

If the data that was missing age information were drawn from the same population as the data with age information, we would expect this last term to be equal to the constant term in the Equation 12. These terms are not the same and we can see from the data that the houses where we do not know the age are substantially leakier than those that we do—all other things being equal.

Regression 8b: Adjustment for floor leakage

The floor leakage variable also had the missing data problem that the age variable had. By repeating the age analysis using the floor leakage variable and using the results to adjust the equation it will be:

$$\ln(NL) = \overleftrightarrow{\beta_{adj.}} + \beta_{Area} \cdot Area + \beta_H \cdot H + \beta_\varepsilon \cdot \varepsilon + \beta_{AT} \cdot AT + \beta_{FL} \cdot FL$$

with: $\overleftrightarrow{\beta_{adj.}} = \overleftrightarrow{\beta_{cz}} - \beta_{AT} \cdot \overline{AT} - \beta_{FL} \cdot \overline{FL}$ **Equation 13**

Regression 8c: Adjustment for Low-income

To determine the low-income parameters we will use just the low-income data with the core model (plus age) and regress:

$$\ln(NL) - (\overleftrightarrow{\beta_{adj}} + \beta_{Area} \cdot Area + \beta_H \cdot H + \beta_\varepsilon \cdot \varepsilon + \beta_{AT} \cdot AT) = \beta_{LI,AT} \cdot AT + \beta_{LI,Area} \cdot Area + \beta_{LI}$$

Equation 14

This determines the last of the parameters we are concerned with. The coefficients describe the difference between the low-income and the ordinary data.

Interpretation of the Results

The full predictive model can be written in the following form:

$$\ln(NL) = \overleftrightarrow{\beta_{adj.}} + \beta_{Area} \cdot Area + \beta_H \cdot H + \beta_\varepsilon \cdot \varepsilon + \beta_{AT} \cdot AT + \beta_{FL} \cdot FL + (\beta_{LI} + \beta_{LI,Area} \cdot Area_{LI} + \beta_{LI,AT} \cdot AT_{LI}) \cdot LI$$

Equation 15

Since, ultimately, we are interested in the Normalized Leakage we manipulate the equation by taking the exponent of both sides. To make the equation more meaningful we create a term, NL_{cz}, which is the Normalized Leakage in a particular climate zone of a building with a floor area of 100 m² and one story high with unknown age, basement type, energy program participation and occupant income. This core is then modified by six parameters, any of which can be dropped if the information is unknown. Three of the parameters (participation in an energy efficiency program, presence of floor leakage, and occupants in a low-income bracket) have an exponent that is a probability so that the average NL can be calculated for a group of houses that have a known distribution of these factors. Equation 16 shows the final version of the predictive model. The values for each of the parameters can be found in Table 6. The predicted values based on the Alaska data only are also included in this table for comparison. The only value that changed significantly was the e-program coefficient. The dataset from the Alaska Housing Finance Corporation used the EPA Energy Star rating system, so a house that was modelled to use 30% less energy than the base case was defined as an e-program house. As we might expect, our model shows that these Energy Star houses also have 30% less leakage that the average house. The e-program houses in our database as a whole generally participated in programs with more stringent requirements, and therefore the leakage of e-program houses in the overall database is lower.

$$NL = NL_{cz} \cdot \phi_{Area}^{size-1} \cdot \phi_{Height}^{N_{story}-1} \cdot \phi_{\varepsilon}^{p_{Eff}} \cdot \phi_{Age}^{Age} \cdot \phi_{Floor}^{p_{Floor}} \cdot \left(\phi_{LI,Age}^{Age} \cdot \phi_{LI,Area}^{size-1} \cdot \phi_{LI}\right)^{P_{LI}}$$

Equation 16

Where $NL_{cz} = e^{\overline{\beta_{adj.}} + \beta_{Area} \cdot Area_{ref} + \beta_H \cdot H_{single\text{-}story}}$ **Equation 17**

$Area_{\mathrm{ref}} = 100m^2$ $H_{\text{single-story}} = 2.5m$

And

$$size = \frac{Area}{Area_{ref}}$$ **Equation 18**

Parameter	Defined as:	**Value (AK only)**	**Value (all data)**
ϕ_{Area}	$\phi_{Area} = e^{\beta_{Area} \cdot Area_{ref}}$	0.867 ± 0.003	0.841 ± 0.003
ϕ_H	$\phi_H = e^{\beta_H \cdot H_{Single-Story}}$	1.158 ± 0.005	1.156 ± 0.005
ϕ_ε	$\phi_\varepsilon = e^{\beta_\varepsilon}$	0.680 ± 0.006	0.598 ± 0.004
ϕ_{Age}	$\phi_{Age} = e^{\beta_{AT}}$	1.0162 ± 0.0002	1.0118 ± 0.0002
ϕ_{Floor}	$\phi_{Floor} = e^{\beta_{FL}}$	n/a	1.08 ± 0.02
ϕ_{LI}	$\phi_{LI} = e^{\beta_{LI}}$	n/a	2.45 ± 0.01
ϕ_{AgeLI}	$\phi_{LI,Age} = e^{\beta_{LI,Age}}$	n/a	0.9942 ± 0.0001
ϕ_{AreaLI}	$\phi_{LI,Area} = e^{\beta_{LI,Area} \cdot Area_{ref}}$	n/a	0.775 ± 0.003
$NL_{CZ(Alaska)}$	Equation 17	0.33 ± 0.01	0.36 ± 0.01
$NL_{CZ(Cold)}$	Equation 17	n/a	0.53 ± 0.01
$NL_{CZ(Humid)}$	Equation 17	n/a	0.35 ± 0.01
$NL_{CZ(Dry)}$	Equation 17	n/a	0.61 ± 0.01

Table 6: Values of parameters for predictive model

Influence of Floor area

The value of 0.84 for the floor area parameter means that for every 100m^2 (1,000ft^2) of floor area added the house's Normalized Leakage gets about 16% lower. It is important to note the limitations of our model here. The Normalized Leakage is normalized by floor area and building height. When we look at the Equivalent Leakage Area (non-Normalized Leakage) predicted by this model we find that the shape of the curve is such that it increases to a point and then decreases at higher areas. Physically, it doesn't make sense for the ELA to decrease with increasing building size so the model shouldn't be used to predict the Normalized Leakage of houses above the inflection point. The regression has set this inflection point at about 400 m^2, and only 1% of our data are larger than this. In the data we see a flat relationship between area and ELA when houses are larger than 400 m^2.

Building Height

The building height parameter, ϕ_H, has a value of 1.16 which means that for a given house the Normalized Leakage increases by 16% if you go from one to two stories, keeping the floor area the same. Because height is one of the parameters used to normalize the leakage it is useful to examine the effect that our model predicts for non-Normalized Leakage. ELA decreases by about 6 % from a 1-story- to a 2-story-building. This makes sense because for the same size of house a two story house will have less surface area, and more of that surface area will be walls.

Leakage often occurs through plumbing and other penetrations through the attic plane, so with smaller attic area we would expect fewer of these penetrations and a relatively tighter building shell. Again, when applying the model to predict Normalized Leakage we need to stay within the range of data used for development of the model. Our dataset contained only ten observations that had more than two stories, thus the predictive model is only applicable to one and two story residences.

Energy Efficient Houses

This result indicates that a house that is part of an energy efficiency program has roughly 60% the leakage of a similar house that is not. For a group of houses, P_{Eff} can be treated as the probability that a house is part of an energy efficiency program. This result shows that the efforts to make buildings tighter have an effect. Both building types (e-program and non e-program) are represented sufficiently in the database. So the conclusion can be applied to the American housing stock.

Testing Age

For non-low income houses, the age-adjustment factor of 1.01 means that houses get on average a bit over 1% leakier every year. Although this effect looks small it can be quite substantial over the life of the house.

Floor leakage

The floor leakage parameter of 1.08 implies that the buildings in this database are 8% leakier if the building has a crawlspace or an unconditioned basement as opposed to a conditioned basement or slab-on-grade construction.

Climate

The variation from climate zone to climate zone is not very big and shown in Table 6. Buildings in the humid climate are the tightest, explained perhaps because houses in the south east tend to be more often built of concrete block which is generally much tighter than wood frame construction. In the other three climates house tightness varies with climate severity as we expect: Alaska houses are the tightest, followed by houses in the cold climate, and houses in the dry climate are the leakiest.

Low Income

For low-income houses, we use the weatherization dataset to modify the coefficients for climate constants and the coefficients for the age and floor area for low-income-buildings. The values, ϕ_{LI}, ϕ_{AgeLI}, and ϕ_{AreaLI} (see Table 6) show how an ordinary building differs from a building with low-income residents.

These three terms are raised to the power of P_{LI} in Equation 16 where P_{LI} is the probability of the building being low-income or, equivalently, the fraction of similar houses that are low-income. The expression for the constant term of 2.45 is a correction of the constant coefficient NL_{cz}. This means houses with low-income residents are about 145% leakier than the same house with non-low income residents.

The low income age coefficient is very slightly less than 1 (0.9942) which indicates that houses with low-income residents become slightly less leaky with age than their counterparts with non-low income residents.

The area coefficient, however, shows a marked difference between houses with low-income residents where the Normalized Leakage decreases by 39% per added $100m^2$ opposed to houses with non-low income residents where the Normalized Leakage decreases only 16% per added $100m^2$. This shows us that the size of a house makes less difference to it's leakage if it is occupied by low income residents than if it is occupied by non low income residents.

Comparison to previous work

The R-squared of the core regression is 30%. This is lower than some of our preliminary regressions, but we have added additional data for this regression. In order to compare this model to the previous model described Chan et. al. [2003] which uses only floor area and year built variables to describe the data we needed to run both models on the same data set. So we re-calculated her parameters (shown in Table 7) based on the new data. The new values are very similar to the values that Chan published. In order to compare the two models we calculated the root mean square of the residuals for Chan's model (147.45) and our model (147.02). The values are so similar that we conclude that both models describe the data equally well.

Type	Coefficient	Estimate	Std. Error	t-value	R-squared
Low-income	(Intercept) Year Built Floor Area	$1.09 \times 10^{+1}$ -5.27×10^{-3} -4.25×10^{-3}	$0.02 \times 10^{+1}$ 0.08×10^{-3} 0.04×10^{-3}	66.1 -62.7 -98.5	0.18
Ordinary	(Intercept) Year Built Floor Area	$1.75 \times 10^{+1}$ -9.12×10^{-3} -1.67×10^{-3}	$0.04 \times 10^{+1}$ 0.18×10^{-3} 0.05×10^{-3}	47.6 -49.5 -35.8	0.14
Energy Program	(Intercept) Year Built Floor Area	$4.13 \times 10^{+1}$ -2.14×10^{-2} -3.22×10^{-4}	$0.15 \times 10^{+1}$ 0.07×10^{-2} 0.60×10^{-4}	27.9 -28.9 -5.3	0.09

Table 7: Coefficients of Chan's Model calculated using the new dataset

Conclusion

Within the scope of this paper a mathematical model to predict the leakage of single-family-buildings was developed. After various regressions to analyze the data a model was fitted which is applicable to the American housing stock within the range of our dataset, defined in Table 8. The model should only be applied to data within these ranges, and it should only be applied to a group of data, and never to just one house.

Variable	# of obs.	Mean	Std. Dev.	Min	Max
Floor Area [m^2]	97222	143	75	31	1035
Height [m]	97512	3.7	1.3	1.7	12.5
Age Tested	79952	35	33	0	370

Table 8: range of the variables in the model

The most significant building characteristic in determining Normalized Leakage was income of the occupants. Buildings where occupants earn less than 125% of the poverty level differ from ordinary buildings in being about 145% leakier. It is important to remember that this conclusion is based on data from only the state of Ohio. We assume that the difference between ordinary and low-income houses in that state can be applied to all other states.

The second most significant characteristic is being part of an energy efficiency program. Buildings that are part of such programs are 40% tighter, on average, than their ordinary counterparts. Although the definition has recently changed, the US EPA Energy Star program has historically defined an energy efficient home as on that is 30% more energy efficient than homes built to the 1993 national Model Energy Code [EPA 2006.] The result shows that the efforts to seal building shells in new construction are successful.

Other significant building characteristics are the building age and floor area of the building. The age of the building has what looks like a small effect of 1% increase in Normalized Leakage per year. But since buildings are used on the order of 100 years the influence grows with the time. This is, in fact, a combined effect of aging and of newer houses having more air tight design and construction materials. As we mentioned, it is impossible, given our data, to separate these effects. The Normalized Leakage decreases by 16% for every additional 100 m^2 of floor area in an ordinary house, and by 39% for each additional 100 m^2 in a house occupied by low-income residents.

When all the exponents in the model are set to zero, the Normalized Leakage is predicted in a particular climate zone for a building with a floor area of 100 m^2 and one story high with unknown age, basement type, energy program participation and occupant income. A more precise prediction can be made if information is available for the floor area, building height, building age, basement type, positive confirmation of participation in an energy efficiency program or the income status of the residents.

The regression method requires that the known variables are random, and that the predicted variable be normally distributed. We know that our data are not randomly sampled, because they come from research programs, weatherization programs and energy rating programs, which all have particular criteria for selecting houses. The predicted variable, the logarithm of Normalized Leakage, is close to normally distributed, but is slightly skewed as we saw in Figure 2. The model could be improved by collecting more data in order to make the database more representative. It would also be possible to weight the representative data and un-weight less representative data for the regression. But in order to do this we need to know which data are representative and which are not.

A next step with this model could be the prediction of Normalized Leakage using U. S. Census Bureau Data to find housing characteristics on a county by county basis. The leakage could then be determined on a county by county basis across the United States.

References

ASHRAE Handbook of Fundamentals, Ch. 27, American Society of Heating, Refrigerating and Air Conditioning Engineers, 2005.

ASHRAE Standard 119, "Air Leakage Performance for Detached Single-Family Residential Buildings", American Society of Heating, Refrigerating and Air Conditioning Engineers, 1988.

Chan, R., P. Price, M. Sohn and A. Gadgil, "Analysis of U.S. Residential Air Leakage Database", LBNL-53367, 2003

Environmental Protection Agency (EPA) Energy Star website, called 03-21-2006
http://www.energystar.gov/index.cfm?c=new_homes.hm_earn_star

Homepage of American Housing Survey, called: 07-19-2005
http://www.census.gov/hhes/www/housing/ahs/ahs03/tab1a5.htm

Homepage of American Housing Survey, called: 07-19-2005
http://www.census.gov/hhes/www/housing/ahs/ahs03/tab1a2.htm

Homepage of Building Science Corporation, called: 07-04-2005
http://www.buildingscience.com/housesthatwork/hygro-thermal.htm

Census of Housing, called: 07-20-2005
http://www.census.gov/hhes/www/housing/census/censushousing.html

Homepage of the Energy Information Administration, called: 07-21-2005
http://www.eia.doe.gov/emeu/reps/maps/us_census_files/cendivco.gif

Homepage of the Energy Information Administration, called: 07-21-2005
http://www.eia.doe.gov/emeu/recs/contents.html

ICC. "International Energy Conservation Code 2003." International Code Council, Country Club Hills, IL., 2003

Little, Roderick J. A. and Rubin, Donald B., Statistical Analysis with Missing Data, Hoboken, NJ: John Wiley & Sons, Inc., 2002

Miller, Irwin and Freund, John E., Probability and Statistics for Engineers, Englewood Cliffs, NJ: Prentice-Hall, Inc., 1985

Sherman, M.H., "Air Infiltration in Buildings", PhD thesis, University of California, Berkeley October 1980

Sherman, M. H., "The Use of Blower-Door Data", LBNL-35173, 1995

Sherman, M. H. and Dickerhoff, D. J., "Air Tightness of U.S. Dwellings",15th Air Infiltration and Ventilation Centre Conference UK, LBNL-48671, 1998

Vining, Geoffrey G., Statistical Methods for Engineers, Pacific Grove, CA: Int. Thomson Publishing, Inc., 1998

White, Frank M., Fluid Mechanics, Hightstown, NJ: McGrawn-Hill, Inc., 1994

Appendices

Appendix A - List of sources for the database

The contributions of the leakage data and the appropriate data were provided by the following organisations and persons:

- Advanced Energy Corporation
- Alaska Housing Finance Corporation
- Arkansas Energy Office
- Building Science Corporation
- Building America
- Building Industry Institute
- Conservation Services Group
- Davis Energy Group
- Rob DeKieffer
- E-Star Colorado
- Geoff Reiler (Sitka, Ak)
- Florida Solar Energy Center
- Guaranteed Wattsavers
- Kansas Energy-Star
- Lawrence Berkeley National Laboratory
- Ohio Home Energy Rating System
- Ohio Weatherization Program
- Vermont Energy Investment Corporation
- Energy Rated Homes of Vermont
- Daran Wastchak, L.L.C.
- Wisconsin Energy Conservation Corporation
- Wisconsin Energy Star Homes

Appendix B - Table with the standardized results of the regressions 1 to 7

This table show the normalized results of regressions 1-7. The results of regression 8 were not normalized. We standardize the variables by subtracting the mean from each value and then dividing by the standard deviation of the distribution. In this way the units of each variable are removed, and the distribution of each variable is centred on zero, with a standard deviation of one.

#	Area	σ	H	σ	AT	σ	YB	σ	FL	σ	DL	σ	e-prog.	σ	LI	σ
	ln(NL) per m^2		ln(NL) per m		ln(NL) per year		ln(NL) per year						ln(NL)		ln(NL)	
1	-0.239	0.003	-0.002	0.003	0.978	0.023	0.8	0.2								
	-0.171	0.004	0.114	0.002	0.527	0.026	0.21	0.03								
2	-0.240	0.002			9.292	0.022	0.77	0.02								
	-0.167	0.003	0.134	0.002	0.085	0.013	-0.22	0.01								
3	-0.161	0.003	0.140	0.002	0.119	0.013	-0.19	0.01	-0.014	0.005	-0.139	0.008				
4	-0.116	0.003	0.088	0.002	-0.037	0.012	-0.29	0.01	-0.007	0.005	-0.174	0.007	-0.216	0.003		
5	-0.185	0.002	0.043	0.001	0.170	0.014	0.00	0.01	-0.012	0.003	-0.076	0.005	-0.158	0.002	0.322	0.004
6	-0.116	0.003	0.085	0.002	0.253	0.004					-0.280	0.007	-0.195	0.003		
	-0.121	0.003	0.091	0.002			-0.290	0.004			-0.284	0.007	-0.205	0.003		
7	-0.118	0.003	0.096	0.002	0.243	0.005					-0.259	0.006	-0.202	0.003		

Appendix C - Table of the non-standardized results of the regressions

This table shows the results of regressions 1-8. The underlined coefficients have a p-test value that is not equal to zero. See Appendix D for P-test and t-test results. The regressions in bold were used in the final result.

#	dataset	description	adj. R^2	obs.	Root MSE	Area	σ	H	σ	AT	σ	YB	σ
		units	%		%	ln(NL) per m^2		ln(NL) per m		ln(NL) per year		ln(NL) per year	
1	ohio	all information known	21.5	50,722	49.82	-0.0044	4.96E-05	-0.003	0.005	0.0349	0.0008	0.029	0.008
	non-o	all information known	36.33	28,908	47.8	-0.0018	3.67E-05	0.130	0.002	0.024	0.001	0.010	0.001
2	ohio	with indicator var.	19.57	50,965	49.88	-0.0044	4.22E-05			0.3312	0.0008	0.0271	0.0008
	non-o	with indicator var.	40.69	43,150	51.39	-0.0020	3.25E-05	0.115	0.002	0.0039	0.0006	-0.0094	0.0006
3	non-o	with floor and duct leakage info	45.72	43,150	49.16	-0.0019	3.14E-05	0.120	0.002	0.0054	0.0006	-0.0080	0.0005
4	non-o	with eprog info	51.38	43,150	46.54	-0.0014	3.07E-05	0.076	0.002	-0.0017	0.0006	-0.0127	0.0005
5	all	together	64.38	94,115	50.24	-0.0025	2.69E-05	0.052	0.002	0.0054	0.0004	0.0001	0.0004
6	non-o	without year built without floor leak.	49.34	43,150	47.50	-0.0014	3.10E-05	0.073	0.002	0.0116	0.0002		
	non-o	without age tested without foor leak.	49.75	43,150	47.32	-0.0014	3.07E-05	0.078	0.002			-0.0125	0.0002
7	non-o	summarized climates	44.96	43,150	49.51	-0.0014	3.25E-05	0.082	0.002	0.0111	0.0002		
8	**core**	**without indicator variables (everyth. known)**	**30.53**	42,874	52.66	**-0.0017**	**3.54E-04**	**0.058**	**0.002**				
a	non-o	to create the age tested coefficient	30.68	28,908	49.88	-0.0016	3.88E-05	0.113	0.002	**0.0117**	**0.0002**		
b	non-o	to create the floor leakage coefficient	49.12	5,646	45.88	-0.0011	7.86E-05	0.102	0.006				
c	**ohio**	**adjusted by low income**	19.23	50,722	50.45	**-0.0026**	**4.28E-05**			**-0.0059**	**0.0001**		

#	dataset	description	FL	σ	DL	σ	e-prog.	σ	LI	σ	climate average	avg. σ
		units					ln(NL)		ln(NL)		ln(NL)	
1	ohio	all information known									-55.5	1.6
	non-o	all information known									-17.9	2.4
2	ohio	with indicator var.									-52.1	1.6
	non-o	with indicator var.									18.4	1.1
3	non-o	with floor and duct leakage info	-0.04	0.01	-0.41	0.02					16.6	2.3
4	non-o	with eprog info	-0.02	0.01	-0.51	0.02	-0.432	0.006			25.5	1.1
5	all	together	-0.05	0.01	-0.34	0.02	-0.422	0.006	0.689	0.009	0.3	0.9
6	non-o	without year built without floor leak.			-0.83	0.02	-0.390	0.006			-0.87	0.02
	non-o	without age tested without foor leak.			-0.84	0.02	-0.410	0.006			24.5	0.4
7	non-o	summarized climates			-0.76	0.02	-0.404	0.006			-0.85	0.02
8	**core**	**without indicator variables (everyth. known)**					**-0.514**	**0.006**			**-0.64**	**0.01**
a	non-o	to create the age tested coefficient					-0.285	0.007			0.00	
b	non-o	to create the floor leakage coefficient	**0.08**	**0.01**			-0.73	0.01			0.00	
c	**ohio**	**adjusted by low income**										

Appendix D – Complete Regression Results

```
      log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression1.log
  log type:  text
 opened on:  14 Mar 2006, 12:09:11
(43150 observations deleted)

      Source |       SS       df       MS              Number of obs =   50722
-------------+------------------------------           F(  5, 50716) = 2779.07
       Model |  3448.78711     5  689.757421           Prob > F      =  0.0000
    Residual |   12587.555 50716   .24819692           R-squared     =  0.2151
-------------+------------------------------           Adj R-squared =  0.2150
       Total |  16036.3421 50721  .316167704           Root MSE      =  .49819

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0043536   .0000496   -87.70   0.000    -.0044509   -.0042563
           h |   -.002629   .0046492    -0.57   0.572    -.0117415    .0064835
  age_tested |   .0348813   .0008195    42.56   0.000      .033275    .0364876
     yrbuilt |   .0288605   .0008135    35.48   0.000      .027266     .030455
         I_c |  -57.22326   1.622429   -35.27   0.000    -60.40324   -54.04328
        I_mh |  -57.04524   1.622432   -35.16   0.000    -60.22523   -53.86526
------------------------------------------------------------------------------

(6009 observations deleted)

      Source |       SS       df       MS              Number of obs =   44855
-------------+------------------------------           F(  4, 44850) = 2988.97
       Model |  2994.91187     4  748.727968           Prob > F      =  0.0000
    Residual |  11234.8065 44850  .250497358           R-squared     =  0.2105
-------------+------------------------------           Adj R-squared =  0.2104
       Total |  14229.7184 44854  .317245248           Root MSE      =   .5005

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0043383   .0000525   -82.66   0.000    -.0044412   -.0042354
           h |  -.0032378   .0049888    -0.65   0.516     -.013016    .0065404
  age_tested |   .0379683   .0008744    43.42   0.000     .0362545     .039682
     yrbuilt |   .0320809   .0008681    36.96   0.000     .0303795    .0337824
         I_c |  -63.63869   1.731292   -36.76   0.000    -67.03205   -60.24533
------------------------------------------------------------------------------
(43150 observations deleted)
(44956 observations deleted)

      Source |       SS       df       MS              Number of obs =    5867
-------------+------------------------------           F(  4,  5862) =  411.31
       Model |   369.51028     4    92.37757           Prob > F      =  0.0000
    Residual |  1316.56742  5862  .224593554           R-squared     =  0.2192
-------------+------------------------------           Adj R-squared =  0.2186
       Total |   1686.0777  5866   .28743227           Root MSE      =  .47391

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
        area |  -.0044507   .0001518   -29.32   0.000    -.0047483   -.0041531
           h |   .0082592   .0126579     0.65   0.514     -.016555    .0330735
  age_tested |   .0107333   .0023182     4.63   0.000     .0061886    .0152779
     yrbuilt |   .0037013   .0022989     1.61   0.107    -.0008053    .0082079
        I_mh |  -6.932204   4.583573    -1.51   0.130     -15.9177     2.05329
------------------------------------------------------------------------------
(50965 observations deleted)

      Source |       SS       df       MS              Number of obs =   28908
-------------+------------------------------           F( 10, 28897) = 1650.43
       Model |  3771.53622    10  377.153622           Prob > F      =  0.0000
    Residual |  6603.51501 28897  .228519051           R-squared     =  0.3635
-------------+------------------------------           Adj R-squared =  0.3633
       Total |  10375.0512 28907  .358911379           Root MSE      =  .47804

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0017819   .0000367   -48.56   0.000    -.0018538     -.00171
           h |    .129987   .0017621    73.77   0.000     .1265332    .1334407
  age_tested |   .0241451   .0012027    20.08   0.000     .0217878    .0265024
     yrbuilt |   .0095019   .0011771     8.07   0.000     .0071948     .011809
        I_sa |  -21.02226   2.353686    -8.93   0.000    -25.63559   -16.40893
        I_vc |  -20.42233   2.353467    -8.68   0.000    -25.03523   -15.80942
         I_c |  -20.40901   2.352278    -8.68   0.000    -25.01958   -15.79843
       I_mhd |  -20.27306   2.351621    -8.62   0.000    -24.88235   -15.66378
        I_mh |   -20.0786   2.350691    -8.54   0.000    -24.68606   -15.47114
        I_hh |  -20.02156   2.350912    -8.52   0.000    -24.62946   -15.41367
         I_m |  -19.75631   2.343056    -8.43   0.000     -24.3488   -15.16381
      I_unkn |  (dropped)
------------------------------------------------------------------------------

 (39414 observations deleted)

      Source |       SS       df       MS              Number of obs =    3052
-------------+------------------------------           F(  4,  3047) =  426.64
       Model |  438.680611     4  109.670153           Prob > F      =  0.0000
    Residual |  783.250573  3047  .257056309           R-squared     =  0.3590
-------------+------------------------------           Adj R-squared =  0.3582
       Total |  1221.93118  3051  .400501863           Root MSE      =  .50701

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0019145   .0001345   -14.23   0.000    -.0021783   -.0016507
           h |   .1170668   .0061106    19.16   0.000     .1050854    .1290482
  age_tested |   .0274868   .0038523     7.14   0.000     .0199334    .0350403
     yrbuilt |  -.0020756   .0036379    -0.57   0.568    -.0092085    .0050574
        I_sa |   2.138296   7.277123     0.29   0.769    -12.13027    16.40686
------------------------------------------------------------------------------
(50965 observations deleted)
(19952 observations deleted)

      Source |       SS       df       MS              Number of obs =   20860
-------------+------------------------------           F(  4, 20855) = 1997.78
       Model |  1849.84083     4  462.460208           Prob > F      =  0.0000
    Residual |  4827.65227 20855  .231486563           R-squared     =  0.2770
```

```
-------------+------------------------------           Adj R-squared =  0.2769
       Total |   6677.4931 20859  .320125275           Root MSE      =  .48113

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0020843   .0000474   -44.01   0.000    -.0021771   -.0019914
           h |   .1350201   .0019723    68.46   0.000     .1311541    .1388861
  age_tested |   .0306998   .0013873    22.13   0.000     .0279807     .033419
     yrbuilt |   .0152002   .0013397    11.35   0.000     .0125742    .0178261
        I_vc |  -31.79712   2.678665   -11.87   0.000    -37.04751   -26.54672
------------------------------------------------------------------------------
(50965 observations deleted)

(32952 observations deleted)

      Source |       SS       df       MS              Number of obs =    3548
-------------+------------------------------           F(  4,  3543) =  653.91
       Model |  456.853345     4  114.213336           Prob > F      =  0.0000
    Residual |  618.823616  3543  .174660913           R-squared     =  0.4247
-------------+------------------------------           Adj R-squared =  0.4241
       Total |  1075.67696  3547  .303263874           Root MSE      =  .41792


------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   -.001123   .0000674   -16.67   0.000     -.001255   -.0009909
           h |   .1150702   .0074686    15.41   0.000     .1004269    .1297134
  age_tested |  -.0433324   .0074556    -5.81   0.000    -.0579501   -.0287148
     yrbuilt |  -.0547123   .0074059    -7.39   0.000    -.0692325   -.0401921
         I_c |   107.8811   14.79541     7.29   0.000     78.87275    136.8895
------------------------------------------------------------------------------
(50965 observations deleted)
(39788 observations deleted)

      Source |       SS       df       MS              Number of obs =     716
-------------+------------------------------           F(  4,   711) =   72.70
       Model |  18.6477544     4  4.66193861           Prob > F      =  0.0000

    Residual |  45.5913897   711  .064122911           R-squared     =  0.2903
-------------+------------------------------           Adj R-squared =  0.2863
       Total |  64.2391441   715  .089844957           Root MSE      =  .25323

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   .0000191   .0001919     0.10   0.921    -.0003577    .0003959
           h |  -.1118513   .0113271    -9.87   0.000    -.1340898   -.0896128
  age_tested |   -.056733   .0073568    -7.71   0.000    -.0711766   -.0422895
     yrbuilt |  -.0609028   .0045984   -13.24   0.000    -.0699309   -.0518747
       I_mhd |   120.8679   9.196544    13.14   0.000     102.8122    138.9235
------------------------------------------------------------------------------
(50965 observations deleted)
(41874 observations deleted)
```

```
      Source |       SS       df       MS              Number of obs =     386
-------------+------------------------------           F(  4,   381) =   79.23
       Model |  48.7605944     4  12.1901486           Prob > F      =  0.0000
    Residual |  58.6185489   381  .153854459           R-squared     =  0.4541
-------------+------------------------------           Adj R-squared =  0.4484
       Total |  107.379143   385  .278906866           Root MSE      =  .39224

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0020813   .0001578   -13.19   0.000    -.0023915   -.0017711
           h |   .2067471   .0176369    11.72   0.000     .1720692     .241425
  age_tested |   .0832178   .0630579     1.32   0.188    -.0407673    .2072029
     yrbuilt |  -.0330728   .0124179    -2.66   0.008     -.057489   -.0086566
        I_mh |   64.76293   24.82295     2.61   0.009     15.95581      113.57
------------------------------------------------------------------------------
(50965 observations deleted)
(42339 observations deleted)

      Source |       SS       df       MS              Number of obs =     325
-------------+------------------------------           F(  4,   320) =   54.37
       Model |  15.2745524     4  3.81863811           Prob > F      =  0.0000
    Residual |   22.474522   320  .070232881           R-squared     =  0.4046
-------------+------------------------------           Adj R-squared =  0.3972
       Total |  37.7490744   324  .116509489           Root MSE      =  .26501

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   .0005094   .0001989     2.56   0.011      .000118    .0009008
           h |   .2674455   .0250376    10.68   0.000     .2181864    .3167046
  age_tested |   .1519625   .0290334     5.23   0.000      .094842     .209083
     yrbuilt |   .0142722   .0065784     2.17   0.031     .0013299    .0272145
        I_hh |  -30.45296    13.1608    -2.31   0.021    -56.34559    -4.56034
------------------------------------------------------------------------------
(50965 observations deleted)
(42857 observations deleted)

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  4,    16) =    2.69
       Model |  .759679726     4  .189919932           Prob > F      =  0.0686
    Residual |  1.12779093    16  .070486933           R-squared     =  0.4025
-------------+------------------------------           Adj R-squared =  0.2531
       Total |  1.88747065    20  .094373533           Root MSE      =  .26549

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0012261   .0010517    -1.17   0.261    -.0034556    .0010034
           h |   .0261567   .0613708     0.43   0.676    -.1039435    .1562569
  age_tested |   .0323883    .085135     0.38   0.709    -.1480898    .2128664
     yrbuilt |    .024657   .0863204     0.29   0.779     -.158334    .2076479
         I_m |  -49.50859   171.6604    -0.29   0.777    -413.4125    314.3953
------------------------------------------------------------------------------
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression1.log
  log type:  text
```

```
 closed on:  14 Mar 2006, 12:09:40
-------------------------------------------------------------------------------
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression2.log
  log type:  text
 opened on:  14 Mar 2006, 12:09:44

-------------------------------------------------------------------------------
-> OWP = 1

      Source |       SS       df       MS              Number of obs =   50965
-------------+------------------------------           F(  6, 50958) = 2308.35
       Model |   3446.3439     6   574.39065           Prob > F      =  0.0000
    Residual |  12679.9541 50958  .248831471           R-squared     =  0.2137
-------------+------------------------------           Adj R-squared =  0.2136
       Total |   16126.298 50964   .31642528           Root MSE      =  .49883

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0043588   .0000422  -103.29   0.000    -.0044415   -.0042761
          AT |   .0331244   .0007959    41.62   0.000     .0315643    .0346844
        I_at |   3.844229   .1679291    22.89   0.000     3.515086    4.173372
          YB |    .027131   .0007901    34.34   0.000     .0255824    .0286796
        I_yb |   50.39544   1.480275    34.04   0.000     47.49408    53.29679
         I_c |   -53.7801   1.575603   -34.13   0.000     -56.8683   -50.69191
        I_mh |  -53.60571   1.575624   -34.02   0.000    -56.69395   -50.51747
------------------------------------------------------------------------------

(50965 observations deleted)

      Source |       SS       df       MS              Number of obs =   43150
-------------+------------------------------           F( 13, 43136) = 2278.28
       Model |  7822.96895    13  601.766842           Prob > F      =  0.0000
    Residual |  11393.5779 43136  .264131534           R-squared     =  0.4071
-------------+------------------------------           Adj R-squared =  0.4069
       Total |  19216.5468 43149  .445353237           Root MSE      =  .51394

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0019596   .0000325   -60.28   0.000    -.0020233   -.0018959
           h |   .1150814   .0016383    70.24   0.000     .1118702    .1182926
          AT |   .0038824   .0006124     6.34   0.000     .0026821    .0050828
        I_at |   .4988667   .0257162    19.40   0.000     .4484624    .5492709
          YB |  -.0094152   .0005686   -16.56   0.000    -.0105298   -.0083007
        I_yb |    -18.708    1.11831   -16.73   0.000    -20.89991   -16.51609
        I_sa |    16.9926   1.136984    14.95   0.000     14.76409    19.22111
        I_vc |   17.54838   1.136923    15.43   0.000     15.31999    19.77677
         I_c |   17.55038   1.136494    15.44   0.000     15.32283    19.77793
       I_mhd |   16.84194   1.136895    14.81   0.000      14.6136    19.07027
        I_mh |   17.74662   1.134337    15.64   0.000     15.52329    19.96994
        I_hh |   17.46194   1.136314    15.37   0.000     15.23474    19.68914
         I_m |    17.2574    1.13638    15.19   0.000     15.03007    19.48472
      I_unkn |   18.05388   1.136812    15.88   0.000     15.82571    20.28205
------------------------------------------------------------------------------
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression2.log
```

```
  log type:  text
 closed on:  14 Mar 2006, 12:09:45
```

```
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression3.log
  log type:  text
 opened on:  14 Mar 2006, 12:09:49
(50965 observations deleted)

      Source |       SS       df       MS              Number of obs =   43150
-------------+------------------------------           F( 17, 43132) = 2139.27
       Model |  8790.73553    17   517.10209           Prob > F      =  0.0000
    Residual |  10425.8113 43132  .241718707           R-squared     =  0.4575
-------------+------------------------------           Adj R-squared =  0.4572
       Total |  19216.5468 43149  .445353237           Root MSE      =  .49165

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0018864   .0000314   -60.00   0.000    -.0019481   -.0018248
           h |   .1199073   .0015742    76.17   0.000     .1168219    .1229927
          AT |   .0054419   .0005903     9.22   0.000     .0042849     .006599
        I_at |   .4745373   .0246488    19.25   0.000     .4262252    .5228495
          YB |  -.0080197   .0005485   -14.62   0.000    -.0090948   -.0069447
        I_yb |   -15.8965   1.078615   -14.74   0.000    -18.01061    -13.7824
          FL |  -.0394595   .0148887    -2.65   0.008    -.0686416   -.0102774
         I_f |  -.2605137   .0145842   -17.86   0.000    -.2890991   -.2319284
          DL |  -.4099247   .0229058   -17.90   0.000    -.4548206   -.3650288
         I_d |   .2905532   .0174694    16.63   0.000     .2563128    .3247937
        I_sa |   14.12794   1.097545    12.87   0.000     11.97673    16.27915
        I_vc |   14.67355    1.09747    13.37   0.000     12.52249    16.82462
         I_c |    14.7045   1.096873    13.41   0.000      12.5546    16.85439
       I_mhd |   14.49288   1.097219    13.21   0.000     12.34231    16.64346
        I_mh |   14.91683   1.095049    13.62   0.000     12.77051    17.06315
        I_hh |   15.13261   1.096763    13.80   0.000     12.98294    17.28229
         I_m |   14.46662   1.096551    13.19   0.000     12.31736    16.61588
      I_unkn |   15.20704    1.09689    13.86   0.000     13.05711    17.35696
------------------------------------------------------------------------------
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression3.log
  log type:  text
 closed on:  14 Mar 2006, 12:09:50
```

```
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression4.log
  log type:  text
 opened on:  14 Mar 2006, 12:09:54
(50965 observations deleted)

      Source |       SS       df       MS              Number of obs =   43150
-------------+------------------------------           F( 18, 43131) = 2533.85
       Model |  9876.61184    18  548.700658           Prob > F      =  0.0000
    Residual |  9339.93498 43131  .216548074           R-squared     =  0.5140
-------------+------------------------------           Adj R-squared =  0.5138
       Total |  19216.5468 43149  .445353237           Root MSE      =  .46535

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
        area |  -.0013535   .0000307   -44.09   0.000    -.0014137   -.0012933
           h |   .0755294   .0016164    46.73   0.000     .0723612    .0786975
          AT |  -.0016929   .0005678    -2.98   0.003    -.0028057   -.0005801
        I_at |   .2117575   .0236235     8.96   0.000     .1654551      .25806
          YB |  -.0126997   .0005233   -24.27   0.000    -.0137254   -.0116739
        I_yb |  -25.02264   1.029015   -24.32   0.000    -27.03953   -23.00575
          FL |  -.0192556   .0140951    -1.37   0.172    -.0468822     .008371
         I_f |  -.3082791   .0138205   -22.31   0.000    -.3353675   -.2811907
          DL |  -.5113452   .0217277   -23.53   0.000    -.5539319   -.4687584
         I_d |   .0124606   .0169948     0.73   0.463    -.0208496    .0457708
       eprog |  -.4321517   .0061027   -70.81   0.000    -.4441132   -.4201903
        I_sa |   24.09063   1.048314    22.98   0.000     22.03591    26.14534
        I_vc |   24.60848   1.048191    23.48   0.000       22.554    26.66295
         I_c |    24.5273   1.047419    23.42   0.000     22.47434    26.58027
       I_mhd |     24.494   1.048081    23.37   0.000     22.43974    26.54826
        I_mh |   24.64762   1.045537    23.57   0.000     22.59835     26.6969
        I_hh |   25.03695    1.04747    23.90   0.000     22.98389    27.09001
         I_m |    24.4073    1.04734    23.30   0.000      22.3545    26.46011
      I_unkn |   25.05098   1.047475    23.92   0.000     22.99791    27.10405
------------------------------------------------------------------------------
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression4.log
  log type:  text
 closed on:  14 Mar 2006, 12:09:55
```

```
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression5.log
  log type:  text
 opened on:  14 Mar 2006, 12:09:59

      Source |       SS       df       MS              Number of obs =   94115
-------------+------------------------------           F( 19, 94095) = 8955.19
       Model |  42953.4689    19  2260.70889           Prob > F      =  0.0000
    Residual |  23753.9869 94095  .252446855           R-squared     =  0.6439
-------------+------------------------------           Adj R-squared =  0.6438
       Total |  66707.4558 94114  .708794183           Root MSE      =  .50244

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0025023   .0000269   -93.09   0.000     -.002555   -.0024496
           h |   .0519382   .0015865    32.74   0.000     .0488288    .0550477
          AT |   .0054197   .0004436    12.22   0.000     .0045502    .0062892
        I_at |   .4610627    .022625    20.38   0.000     .4167179    .5054075
          YB |   .0001015   .0004353     0.23   0.816    -.0007517    .0009547
        I_yb |   .1757688   .8550894     0.21   0.837    -1.500197    1.851735
          FL |  -.0543379   .0149671    -3.63   0.000    -.0836732   -.0250025
         I_f |  -.3626044   .0141717   -25.59   0.000    -.3903807   -.3348281
          DL |  -.3443546   .0229151   -15.03   0.000    -.3892681   -.2994412
         I_d |    .126643   .0181193     6.99   0.000     .0911293    .1621567
       eprog |  -.4218036   .0063925   -65.98   0.000    -.4343329   -.4092743
         OWP |   .6893771   .0085315    80.80   0.000     .6726555    .7060987
        I_sa |   -1.24745   .8716789    -1.43   0.152    -2.955931    .4610316
        I_vc |  -.7280879    .871558    -0.84   0.404    -2.436332    .9801562
         I_c |  -.6716203   .8707128    -0.77   0.441    -2.378208    1.034967
       I_mhd |  -.8473643    .871601    -0.97   0.331    -2.555693    .8609642
```

```
        I_mh |   -.498568   .8705385    -0.57   0.567    -2.204814    1.207678
        I_hh |  -.2702527   .8711514    -0.31   0.756      -1.9777    1.437195
         I_m |  -.8122206   .8716544    -0.93   0.351    -2.520654    .8962126
      I_unkn |  -.2176131   .8709985    -0.25   0.803    -1.924761    1.489535
------------------------------------------------------------------------------
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression5.log
  log type:  text
 closed on:  14 Mar 2006, 12:10:00
```

```
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression6.log
  log type:  text
 opened on:  14 Mar 2006, 12:10:04
(50965 observations deleted)

      Source |       SS       df       MS              Number of obs =   43150
-------------+------------------------------           F( 14, 43135) = 3002.45
       Model |  9484.09625    14  677.435447           Prob > F      =  0.0000
    Residual |  9732.45056 43135  .225627694           R-squared     =  0.4935
-------------+------------------------------           Adj R-squared =  0.4934
       Total |  19216.5468 43149  .445353237           Root MSE      =    .475

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   -.001354    .000031   -43.63   0.000    -.0014148   -.0012931
           h |   .0730916   .0016453    44.43   0.000     .0698669    .0763164
          AT |   .0115873   .0002002    57.87   0.000     .0111948    .0119797
        I_at |   .7098758   .0070239   101.07   0.000     .6961088    .7236429
          DL |   -.825268   .0192796   -42.81   0.000    -.8630564   -.7874797
         I_d |  -.2057197   .0153738   -13.38   0.000    -.2358527   -.1755868
       eprog |   -.389691   .0061196   -63.68   0.000    -.4016855   -.3776966
        I_sa |  -1.392192    .020066   -69.38   0.000    -1.431521   -1.352862
        I_vc |   -.880287   .0190594   -46.19   0.000    -.9176438   -.8429301
         I_c |   -.990322   .0175133   -56.55   0.000    -1.024648   -.9559955
       I_mhd |  -1.054638   .0210255   -50.16   0.000    -1.095849   -1.013428
        I_mh |  -.6636532    .019602   -33.86   0.000    -.7020734   -.6252329
        I_hh |   -.458584   .0255108   -17.98   0.000    -.5085858   -.4085823
         I_m |  -1.139075   .0331251   -34.39   0.000    -1.204001    -1.07415
      I_unkn |  -.5492496   .0337859   -16.26   0.000    -.6154706   -.4830286
------------------------------------------------------------------------------


      Source |       SS       df       MS              Number of obs =   43150
-------------+------------------------------           F( 14, 43135) = 3048.21
       Model |  9556.76745    14  682.626246           Prob > F      =  0.0000
    Residual |  9659.77937 43135  .223942955           R-squared     =  0.4973
-------------+------------------------------           Adj R-squared =  0.4972
       Total |  19216.5468 43149  .445353237           Root MSE      =  .47323

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0014213   .0000307   -46.32   0.000    -.0014815   -.0013612
           h |   .0778014   .0016342    47.61   0.000     .0745983    .0810045
          YB |  -.0125191   .0001788   -70.03   0.000    -.0128695   -.0121687
```

```
        I_yb |  -24.34148   .3549566   -68.58   0.000     -25.0372   -23.64576
          DL |  -.8378601   .0193099   -43.39   0.000    -.8757079   -.8000122
         I_d |  -.2182162   .0153402   -14.23   0.000    -.2482833   -.1881491
       eprog |   -.410248   .0060629   -67.67   0.000    -.4221314   -.3983646
        I_sa |   23.66479   .3573766    66.22   0.000     22.96433    24.36526
        I_vc |   24.17206   .3576374    67.59   0.000     23.47108    24.87303
         I_c |    24.0717   .3550929    67.79   0.000     23.37571    24.76769
       I_mhd |   24.03704   .3568936    67.35   0.000     23.33752    24.73656
        I_mh |   24.40183   .3551675    68.71   0.000      23.7057    25.09797
        I_hh |   24.59992   .3576326    68.79   0.000     23.89896    25.30089
         I_m |   24.13155   .3541066    68.15   0.000     23.43749     24.8256
      I_unkn |   24.50218   .3570192    68.63   0.000     23.80242    25.20195
------------------------------------------------------------------------------
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression6.log
  log type:  text
 closed on:  14 Mar 2006, 12:10:05
```

```
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression7.log
  log type:  text
 opened on:  14 Mar 2006, 12:10:09
(50965 observations deleted)

      Source |       SS       df       MS              Number of obs =   43150
-------------+------------------------------           F( 11, 43138) = 3204.65
       Model |  8641.57304    11  785.597549           Prob > F      =  0.0000
    Residual |  10574.9738 43138  .245142885           R-squared     =  0.4497
-------------+------------------------------           Adj R-squared =  0.4496
       Total |  19216.5468 43149  .445353237           Root MSE      =  .49512

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0013796   .0000325   -42.48   0.000    -.0014432   -.0013159
           h |   .0821337   .0017073    48.11   0.000     .0787875      .08548
          AT |   .0111427   .0002091    53.28   0.000     .0107328    .0115526
        I_at |    .677965   .0072154    93.96   0.000     .6638227    .6921072
          DL |  -.7644717   .0185183   -41.28   0.000     -.800768   -.7281755
         I_d |  -.1735259   .0157953   -10.99   0.000    -.2044849   -.1425668
       eprog |  -.4038536   .0063232   -63.87   0.000    -.4162471     -.39146
     I_humid |  -.6185201   .0195148   -31.69   0.000    -.6567694   -.5802708
       I_dry |  -1.097901   .0208282   -52.71   0.000    -1.138725   -1.057077
    I_alaska |  -1.009366    .019543   -51.65   0.000    -1.047671   -.9710616
      I_cold |  -1.025855   .0181341   -56.57   0.000    -1.061399   -.9903124
      I_unkn |  -.5857182   .0351498   -16.66   0.000    -.6546126   -.5168239
------------------------------------------------------------------------------
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression7.log
  log type:  text
 closed on:  14 Mar 2006, 12:10:09
```

```
       log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression8.log
  log type:  text
 opened on:  14 Mar 2006, 12:10:14
(276 observations deleted)
```

```
-> non0 = 1

      Source |       SS       df       MS              Number of obs =   42874
-------------+------------------------------           F(  6, 42867) = 3142.68
       Model |  5740.04571     6  956.674285           Prob > F      =  0.0000
    Residual |  13049.3066 42867  .304413807           R-squared     =  0.3055
-------------+------------------------------           Adj R-squared =  0.3054
       Total |  18789.3524 42873  .438256067           Root MSE      =  .55174

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0017372   .0000354   -49.12   0.000    -.0018065   -.0016679
           h |    .058066   .0018152    31.99   0.000     .0545081    .0616238
       eprog |  -.5143286   .0064425   -79.83   0.000     -.526956   -.5017012
     I_humid |  -.3560121   .0146004   -24.38   0.000    -.3846291   -.3273951
       I_dry |  -.8943106   .0125962   -71.00   0.000    -.9189994   -.8696217
    I_alaska |  -.8730227   .0100414   -86.94   0.000    -.8927041   -.8533413
      I_cold |    -.49782   .0102282   -48.67   0.000    -.5178675   -.4777725
------------------------------------------------------------------------------


    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  age_tested |   28908    5.548914    15.38079          0        170

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          DL |    4996    .7345877    .4415967          0          1

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          FL |    5646    .6710946    .4590827          0          1
```

```
-> non0 = 1

      Source |       SS       df       MS              Number of obs =   28908
-------------+------------------------------           F(  7, 28900) = 1829.02
       Model |  3185.20489     7  455.029269           Prob > F      =  0.0000
    Residual |  7189.84634 28900   .24878361           R-squared     =  0.3070
-------------+------------------------------           Adj R-squared =  0.3068
       Total |  10375.0512 28907  .358911379           Root MSE      =  .49878

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0016008   .0000388   -41.26   0.000    -.0016769   -.0015248
           h |   .1130956   .0019851    56.97   0.000     .1092047    .1169865
       eprog |  -.2852972   .0073955   -38.58   0.000    -.2997927   -.2708017
  age_tested |   .0117431   .0002251    52.17   0.000     .0113019    .0121842
     I_humid |  -.8288597   .0220569   -37.58   0.000    -.8720923   -.7856271
       I_dry |  -1.055095   .0210624   -50.09   0.000    -1.096379   -1.013812
    I_alaska |  -1.350404   .0115028  -117.40   0.000     -1.37295   -1.327858
      I_cold |  -1.234993   .0156512   -78.91   0.000     -1.26567   -1.204316
------------------------------------------------------------------------------
```

```
      Source |       SS       df       MS              Number of obs =    4996
-------------+------------------------------           F(  6,  4989) =  724.42
       Model |  593.31362     6  98.8856034            Prob > F      =  0.0000
    Residual | 681.018109  4989   .13650393            R-squared     =  0.4656
-------------+------------------------------           Adj R-squared =  0.4649
       Total | 1274.33173  4995  .255121467            Root MSE      =  .36946

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0010278   .0000695   -14.79   0.000    -.0011641   -.0008915
           h |   .0760192   .0055779    13.63   0.000     .0650841    .0869543
       eprog |  -.5824843   .0160669   -36.25   0.000    -.6139824   -.5509862
          DL |  -.1470902   .0157145    -9.36   0.000    -.1778976   -.1162829
     I_humid |  -.5684585   .0281667   -20.18   0.000    -.6236776   -.5132395
       I_dry |  -.9951699   .0294188   -33.83   0.000    -1.052844   -.9374961
    I_alaska |  (dropped)
      I_cold |  -.7514015   .0304725   -24.66   0.000     -.811141   -.6916621
------------------------------------------------------------------------------

      Source |       SS       df       MS              Number of obs =    5646
-------------+------------------------------           F(  7,  5638) =  779.54
       Model | 1148.66866     7  164.095523            Prob > F      =  0.0000
    Residual | 1186.81459  5638  .210502765            R-squared     =  0.4918
-------------+------------------------------           Adj R-squared =  0.4912
       Total | 2335.48325  5645  .413725997            Root MSE      =  .45881

------------------------------------------------------------------------------
        lnNL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0011218   .0000786   -14.28   0.000    -.0012758   -.0009677
           h |   .1019827   .0057688    17.68   0.000     .0906735    .1132918
       eprog |  -.7327949   .0132403   -55.35   0.000    -.7587509   -.7068389
          FL |   .0756503   .0144094     5.25   0.000     .0474025    .1038982
     I_humid |  -.4126376   .0250786   -16.45   0.000    -.4618013    -.363474
       I_dry |  -.6279918   .2058411    -3.05   0.002     -1.03152    -.224464
    I_alaska |  -.1914362   .0250512    -7.64   0.000    -.2405463   -.1423262
      I_cold |  -.7946466   .0283261   -28.05   0.000    -.8501767   -.7391164
------------------------------------------------------------------------------

-> Oh = 1

      Source |       SS       df       MS              Number of obs =   50722
-------------+------------------------------           F(  2, 50719) = 6040.79
       Model | 3074.46473     2  1537.23236            Prob > F      =  0.0000
    Residual |  12906.742 50719  .254475482            R-squared     =  0.1924
-------------+------------------------------           Adj R-squared =  0.1923
       Total | 15981.2067 50721  .315080671            Root MSE      =  .50446

------------------------------------------------------------------------------
         dif |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |  -.0025546   .0000428   -59.73   0.000    -.0026384   -.0024708
  age_tested |  -.0058538   .0000838   -69.87   0.000     -.006018   -.0056896
       _cons |   .8291634    .005906   140.39   0.000     .8175876    .8407392
------------------------------------------------------------------------------
```

```
      log:
Y:\Residential\Leakage_Database\2005_melanie\new_analysis_2006\Regression8.log
  log type:  text
 closed on:  14 Mar 2006, 12:10:18
```