Causal Factors and Costs of Home Plumbing Corrosion: An Investigation of Sample Selection Bias

Ewa J. Kleczyk Department of Agricultural & Applied Economics Virginia Tech, VA. 24061 TargetRx Horsham, PA. 19044 Phone: (215) 444-8806 Email: ewak@vt.edu	**Darrell J. Bosch** Department of Agricultural & Applied Economics Virginia Tech, VA. 24061 Phone: (540) 231-5265 Email: bosch@vt.edu

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Long Beach, California, July 23-26, 2006

Causal Factors and Costs of Home Plumbing Corrosion: An Investigation of Sample Selection Bias

Ewa J. Kleczyk and Darrell J. Bosch[1]

Abstract

High incidences of pinhole leaks, which occur in home plumbing due to pitting corrosion of water pipes, have been observed in parts of the U.S. such as the Maryland suburbs of Washington D.C. This research evaluates factors associated with pinhole leak occurrences and assesses the costs incurred by consumers from corrosion. Statistical analysis of Maryland survey responses suggests that the probability of pinhole leak occurrences is associated with the type of pipes installed and the distance of the dwelling from a water treatment plant. The number of leaks and location of pinhole leaks in the dwelling drive the financial costs of pinhole leak damage. Correcting for sample selection bias influences the estimated coefficients and statistical significance of the model. Research findings will inform policymakers, program managers, and water utilities on the importance of reducing corrosion in home drinking water infrastructure.

Introduction

Pinhole leaks are small holes in drinking water plumbing caused by pitting corrosion, a type of corrosion concentrated on a very small area of the pipe. Theories as to the origins of pinhole leaks vary and there are a few proven causes (Edwards et. al, 2004). Edwards suggests that removal of natural organic matter mandated by tighter Environmental Protection Agency (EPA) drinking water standards can contribute to the problem in combination with other factors, since natural organic matter can be an inhibitor to the corrosion-inducing chemical reactions. Pipe failures can also result from other factors including faulty installation or wearing of copper due to friction by rubbing against a surface (Fleishman, 2001).

[1] Doctoral student, Department of Agricultural and Applied Economics, Virginia Tech and Statistical Analyst, TargetRx; Professor, Department of Agricultural and Applied Economics, Virginia Tech. Funding for this research was provided by the National Science Foundation, Award # DMI-0329474. The views expressed in this report are those of the authors and not of the National Science Foundation. Marc Edwards and Robert Buglass provided helpful comments on an earlier version of the paper.

In most cases, very small pinhole leaks are hard to detect, especially, if they appear in pipes running through walls and ceilings. Generally, it is recommended that pipes be replaced after three or four leaks (Gurner, 2003), although others have recommended two leaks (Edwards et al., 2004). Damage from pinhole leaks can include collapse of walls and ceilings and the water can contribute to growth of mold on the surface of walls, floors, and ceilings. Mold exposure can cause allergic reactions, such as irritation of eyes, skin, and throat. Furthermore, copper itself can cause severe health problems such as liver and kidney failure, if consumed in doses higher than 1.3 mg/l (EPA, 1992).

Pinhole leaks appear to be a nationwide problem, but they are more common in certain regions of selected states including Maryland, California, Florida, and Ohio. Some areas (for example, parts of Florida) have considered banning the use of copper piping in order to control the rising number of incidents (Gurner, 2003).

The objectives of this research are 1) to evaluate the frequency and potential causal factors associated with pinhole leaks in home plumbing; and 2) to assess the financial, time, and stress costs incurred by consumers and their home insurers in repairing damages to plumbing and property resulting from pinhole leaks. Descriptive statistics of survey data and regression analysis are employed to assess pinhole leak occurrences in copper plumbing system and costs associated with the leaks. The influence of sample selection bias on estimated coefficients and statistical significance of the cost model is assessed.

Maryland Drinking Water Assessment Survey

In July 2004, the Maryland Home Drinking Water Assessment Survey was conducted to learn more about the extent of pinhole leak problems in household drinking water plumbing systems. Mail surveys were sent to Maryland residents in the suburbs of Washington, D.C. to investigate their experiences with pinhole leaks. This area was selected because of the large number of pinhole leaks reported by utility customers to the Washington Suburban Sanitary Commission (WSSC). The sample design was intended to fulfill two purposes[2]: 1) provide a gauge of the incidence of corrosion and pinhole leak problems across the study area; and 2) target areas with high reported rates of pinhole leaks in order to provide more information about pinhole leaks in residential dwellings. The sample was divided by zip code and zip codes with high numbers of reported pinhole leaks were sampled more heavily. A minimum of 10 surveys was sent to every zip code reporting leaks.

A total of 5,009 Maryland residents received the survey and 1,128 responses were returned of which 1,120 responses were used in the analysis. Eight responses were dropped from the study because they were incomplete. The survey analysis focused on the incidence of pinhole leaks; associated financial, time, and emotional costs; and potential causal factors. Samples were weighted to correct for oversampling in zip codes with high numbers of reported pinhole leaks. Responses were weighted by the number of people over 18 represented by each survey sample in each zip code. Weighting was done as follows: 1) *Calculate the number of residents over age 18 per sample survey in the study area*. There were an average of 216 residents per sample survey (1,083,323 residents divided by 5,009 surveys). 2) *Calculate the number of residents over age 18 per sample survey in an individual zip code area.* For example, a zip code area with a population of 35,000, which received 100 surveys would have 350 residents per sample. 3) *Construct the weight for an individual zip code area as the ratio of the number of residents per sample in the individual study area to the number of residents per*

[2] Advice on sampling protocol was provided by Dr. Robert Schulman and Ying Zhang of the Virginia Tech, Statistical Consulting Center, April 2004.

sample in the overall study area. Responses from the zip code example would have a weight of 1.6 (= 350/216).

Pinhole Leak Occurrences

Descriptive Statistics

Survey results indicate that pinhole leaks in copper home plumbing systems are a frequently reported problem in the Maryland suburbs of Washington, D.C. Four hundred forty-eight respondents (unweighted basis) reported incidents of pinhole leaks. Most respondents (74 percent on a weighted basis) had one to four leaks; however, 8 percent reported seven or more leaks (Table 1). Forty-three percent of respondents (weighted basis) reported that the first leak occurred since 2000, 38 percent reported the first leak in the 1990's, 6 percent reported the first leak in the 1980's, and less than 1 percent reported the first leak before 1980 (Table 2).

Table 1. Number of Leaks Reported by Survey Respondents

Number of leaks	Respondents by leak category (unweighted)	Percent (unweighted)	Respondents by leak category (weighted)[a]	Percent (weighted)[a]
1-2	201	44.9	172.5	58.5
3-4	97	21.7	45.9	15.6
5-6	49	10.9	20.4	6.9
7-10	37	8.3	17.5	5.9
> 10	25	5.6	6.9	2.4
Don't know	21	4.7	16.3	5.5
Missing	18	4.0	15.2	5.2
Total	448	100.0	294.7	100.0

[a]Results are weighted based on sampling weights.

Table 2: The Year Pinhole Leaks Were First Reported				
Year pinhole leaks were first reported	Respondents by year leaks first reported (unweighted)	Percent (unweighted)	Respondents by year leaks first reported (weighted)[a]	Percent (weighted)[a]
Since 2000	204	45.5	126.4	42.9
1995 to 1999	124	27.7	75.1	25.5
1990 to 1994	50	11.2	37.3	12.6
1980 to 1989	29	6.5	18.1	6.1
1970 to 1979	7	1.6	1.5	0.5
1960 to 1969	1	0.2	0.1	0.0
Before 1960	1	0.2	0.5	0.2
Do not Know	14	3.1	19.7	6.7
Missing observation	18	4.0	16.0	5.4
Total	448	100.0	294.7	100.0
[a]Results are weighted based on sampling weights.				

According to the survey, the age distribution of dwellings with leaks is roughly proportional to the overall age distribution of dwellings. For example, on an unweighted basis (Table 3), dwellings built prior to 1970 account for 65 percent of the total observations and 63 percent of the leaks. The largest discrepancy is in the 1960-1969 period. Dwellings from that period account for 20 percent of the total observations but 31 percent of the leaks (unweighted basis).

Most leaks were located in the basement (61 percent on a weighted basis) followed by the first floor (37 percent) (Table 4). Only 19 percent of respondents (weighted basis) reported leaks occurring on the second floor or higher.

Table 3: Age of the Dwelling in which Pinhole Leaks Occurred

Year house or apartment building was built	Total number of respondents by year residence built (unweighted)	Percent (unweighted)	Respondents with leaks by year residence built (unweighted)	Percent (unweighted)	Respondents with leaks by year residence built (weighted)[a]	Percent (weighted)[a]
Since 2000	26	2.3	1	0.2	0.1	0
1990 to 1999	52	4.6	8	1.8	8.7	2.9
1980 to 1989	136	12.1	52	11.6	52.1	17.7
1970 to 1979	124	11.1	47	10.5	45.5	15.4
1960 to 1969	225	20.1	82	31.3	59.9	20.3
1950 to 1959	273	24.4	140	15.4	75.8	25.7
1940 to 1949	135	12.1	69	8.3	34.3	11.6
Before 1940	93	8.3	37	8.3	9.8	3.3
Do not know	51	4.6	10	2.2	8.5	2.9
Missing observations	5	0.4	2	0.4	.2	0.1
Total	1120	100.0	448	100.0	294.7	100.0

[a] Results are weighted based on sampling weights

Table 4. Level of dwelling on which leaks occurred.

Level of leak in dwelling	Respondents with leaks at indicated level (unweighted)	Percent[a] (unweighted)	Respondents with leaks at indicated level (weighted)	Percent[a] (weighted)
Under slab /underground	16	3.6	18.3	6.2
Basement	304	67.9	179.8	61.0
First floor	173	38.6	109.3	37.1
Second floor	95	21.2	3.9	18.3
Third floor	8	1.8	0.8	0.3
Fourth floor or higher	7	1.6	1.4	0.5
Don't know	17	3.8	29.3	9.9
Missing	18	4.0	16.1	5.5

[a] Percent calculated based on the number of respondents who indicated the category divided by the total number of respondents with leaks (448 unweighted, 294.7 weighted). Percents do not sum to 100 because some respondents indicated more than one level of dwelling had leaks.

Regression Analysis of Pinhole Leak Occurrences

A probabilistic analysis of potential causal factors using a weighted logistic model was performed to predict the probability of pinhole leaks occurrences. The logistic regression predicts the probability of pinhole leak incidents associated with the independent factors. The probabilities of pinhole leak occurrences are presented as follows:

$$P_i = \Phi(z_i{}'\boldsymbol{\beta}) \quad (1),$$

where $\Phi(z'_i\boldsymbol{\beta})$ is the cumulative distribution of a logistic random variable, z'_i explanatory variables, and $i = 1 \ldots n$ (Griffiths et al., 1993). The cumulative distribution function is provided by

$$P_i = \Phi(z_i{}'\boldsymbol{\beta}) = 1/(1 + e^{(-zi'\boldsymbol{\beta})}) \quad (2).$$

Logistic regression does not depend on the restrictive assumptions of least square regression such as linear relationship between dependent and independent variables, normally distributed endogenous variables and error terms, and homogeneity of variance (Garson, 2006). The logistic probability density function is smooth, symmetric around zero, and bell-shaped. Maximum likelihood estimation is employed in logistic function evaluation (Griffiths et al., 1993).

To evaluate the logistic regression, the explanatory variables are related to the probability of pinhole leak incidents. The dependent variable is binary taking a value of 1 when pinhole leaks occur and 0 otherwise. The following causal factors were employed in the model: geographical location of dwelling relative to water treatment plant (approximated by estimated water travel time to zip code in which the dwelling is located)[3], type of pipes installed (copper vs. others), history of pipe failure, time of pipe replacement (1960-1980), age of the dwelling, and the source of water (approximated by a dummy variable with a value of 1 for respondents

[3] Estimated water travel times were provided by Robert Buglass, Principal Environmental Engineer, Washington Suburban Sanitary Commission, Laurel, Maryland. March, 2006.

located in zip codes which predominantly receive their water from Patuxent water treatment plant and 0 otherwise).[4] History of pipe failure reflects whether the respondent had other types of failure in their drinking water pipes beside pinhole leaks. Time of pipe replacement is a dummy variable with a value of 1 for respondents who indicated they had replaced or installed new drinking water pipes between 1960 and 1980. Age of the dwelling variable is approximated by a dummy variable with value of 1 for dwellings built before 1970 and value of 0 otherwise.

The total number of unweighted observations with complete information on all variables utilized in the regression estimation was 879 (423 respondents with pinhole leaks and 456 respondents without pinhole leaks). The total unweighted number of observations with missing values is 241. Respondents without pinhole leaks were the control group in the logistic regression analysis. The logistic model is significant at a 5 percent level with Log Likelihood equal to 1,170 (Table 5, Base model results).

The logistic regression results are presented in form of odds ratio and probability values (Table 5). The estimated odds ratios indicate relationships between the dependent and independent variables. When the odds ratio is more than 1, there is a positive relationship between the dependent and independent variables and pinhole leaks are more likely to occur with a one unit increase in the value of the explanatory factor. An odds ratio less than 1 indicates a negative relationship between the causal factor and pinhole leaks. For example, an odds ratio of 0.59 for water travel time indicates that pinhole leaks are about half as likely to occur with a one-day increase in travel time. Finally, an odds ratio of 1 presents no relationships between leaks and its causes (UCLA Academic Technology Services, 2005). The estimated probability coefficients (Table 5) can be interpreted as the impact of one unit increase (decrease) in the independent variable on the chance of pinhole leak occurrences, while controlling for other

[4] Almost all respondents received water from WSSC.

variables in the model (UCLA Academic Technology Services, 2005). For example, one additional day of water travel decreases the probability of pinhole leaks by 37 percent. This finding is in agreement with Rushing and Edwards (2004) who found houses located closer to the treatment plant to be exposed to higher levels of chlorine and, therefore, to experience higher levels of pipe corrosion. Furthermore, copper pipes installed in the dwelling raise the chance of pinhole leak incidents by 66 percent; and history of pipe failures increases the probability of pinhole leaks by 69 percent. Pipe replacement, age of dwelling, and the source of water are not statistically significant.

Table 5: Effects of Selected Variables on Probability of Pinhole Leak Occurrences[a]

	Base model[a]			Observations weighted by age category of dwelling[b]		
Variables	Odds Ratio	Probability	P-value	Odds Ratio	Probability	P-value
Water travel time	0.59	-36.94	0.00[c]	0.58	-36.52	0.00[c]
Copper pipes	1.84	64.85	0.00[c]	1.74	63.51	0.00[c]
History of pipe failure	2.25	69.23	0.00[c]	2.15	68.25	0.00[c]
Pipe replaced between 1960 – 1980	0.94	-48.35	0.86	0.94	-48.55	0.88
Dwelling built before 1970	1.10	52.46	0.47	1.00	50.06	0.01[c]
Source of water (Patuxent treatment plant)	1.06	51.40	0.74	1.09	52.24	0.59

[a] Dependent and independent variables were weighted by sample weights.

[b] Observations weighted by sample weights and age of dwelling as described in text.

[c] Significant at 5%.

[d] Base model: Log Likelihood = 1170; Pseudo R^2 = 0.053; n (unweighted) = 879.

[e] Observations weighted by age category of dwelling model: Log Likelihood = 1168 ; Pseudo R^2 = 0.06; n (unweighted) = 879.

A prediction accuracy analysis was performed to determine correct and incorrect estimates of the pinhole leak observations. Correct estimates are defined those responses for whom a probability greater than 0.5 is assigned to a respondent with pinhole leaks or a probability less than 0.5 is assigned to a respondent with no reported pinhole leaks. The analysis does not reveal, however, how close to 1.0 the correct predictions are nor how close to 0.0 the errors are (Garson, 2006). According to the analysis (Table 6), 60 percent of the unweighted values with or without pinhole leaks were correctly classified. Furthermore, 239 observations classified as no pinhole leaks should have been classified as observations with pinhole leaks, while 309 observations with pinhole leaks were predicted appropriately. As a result, the predicted number of pinhole leaks observations is 548 compared to the observed number of observations with pinhole leaks of 423. In a perfect model, all cases are on the diagonal and the percent of correctly predicted values is 100%.

Table 6: Classification Table of Pinhole Leak Prediction Accuracy[a,b,c]

Observed	Predicted		Percentage Correct
	No pinhole leaks	Pinhole leaks	
No pinhole leaks	217	239	48
Pinhole leaks	114	309	73
Overall Percentage			60

[a] Initial Log-likelihood = 1219.
[b] The cutoff probability value is .500.
[c] Values shown are unweighted.

Costs Associated with Pinhole Leaks

Descriptive Statistics

Households with pinhole leaks incur high financial, time, and emotional costs when dealing with pipe failures. One third of respondents (weighted data) with leaks reported expenditures of at least $500 for repairing leaks and associated damage (Table 7). Seven respondents cited $12,000 in repair expenses and one person reported more than $25,000 in damage. Many of the survey respondents pointed out that in addition to repairing or replacing water pipes, they also had to repair damage to ceilings, walls, and floors. In addition, a few respondents moved out of their homes during the renovation periods, which added to their repair expenses. Others mentioned the priceless worth of personal belongings such as family photos, clothes, and furniture damaged by pipe leaks.

Table 7: Costs of Repairing Pinhole Leaks and Associated Damage

Amount of money	Number of observations (unweighted)	Percent (unweighted)	Number of observations (weighted)	Percent (weighted)
Less than $100	58	12.9	55.1	18.7
$100 to $500	112	25.0	89.2	30.3
$501 to $1,000	71	15.8	33.7	11.4
$1,001 to $3,000	66	14.7	32.6	11.1
$3,001 to $5,000	45	10.0	22.5	7.6
More than $5,000	43	9.6	14.6	4.9
Do not know	34	7.6	29.7	10.1
Missing observations	19	4.2	17.4	5.9
Total	448	100.0	294.7	100.0

Twenty-six percent of respondents with pinhole leaks (Table 8, weighted results) spent at least 20 hours dealing with pinhole leaks and related damage. Forty-three percent of respondents spent less than 10 hours. Fifty-six percent of respondents found the experience somewhat or very stressful. Survey respondents felt “aggravated,” “helpless,” and “worried” about future leaks. Lack of knowledge on the possible causes of pinhole leaks added to the overall anxiety felt by those with pinhole leaks. Furthermore, some respondents cited lack of third party involvement and responsibility for pinhole leak occurrences and related damage as additional stress. The respondents felt they had inadequate information to make decisions on water plumbing issues and were frustrated with insufficient assistance from local water utilities, insurance companies, and contractors.

Table 8: Time Spent Dealing with Pinhole Leak Problems				
Number of hours	Number of respondents (unweighted)	Percent (unweighted)	Number of respondents (weighted)	Percent (weighted)
Less than 10	173	38.6	125.4	42.6
11-20	88	19.6	46.2	15.7
21-40	60	13.4	31.0	10.5
41-80	45	10.0	33.5	11.4
More than 80	26	5.8	11.5	3.9
Do not know	36	8.0	29.8	10.1
Missing observations	20	4.5	17.2	5.8
Total	448	100	294.7	100

Regression Analysis of Costs Associated with Pinhole Leak Repair

To assess the factors driving costs of pinhole leak repairs and associated damage, a weighted ordinary least squares (OLS) regression was estimated. The ordinary least square regression assumes i) error term (e_i) is independently and identically distributed in i with a standard normal distribution; ii) $\{x_i: i,\ldots,n\}$ is independent of e_i; iii) error term (e_i) and unmeasured latent continuous random variable (δ_i) are independent within respondent i. δ_i is distributed normally and represents the unmeasured characteristics of respondents (Briggs, 2002). The weighted ordinary least squares regression is specified as follows:

$$y = X\boldsymbol{\beta} + \boldsymbol{e} \qquad (3)$$

$$\boldsymbol{e} \sim (0, \sigma^2 I_i) \qquad (4)$$

where y is the dependent variable weighted by observation, and X is a vector of independent variables weighted by observation (Briggs, 2004).

The dependent cost variable (y) includes the financial and time costs of pinhole leaks. Time spent dealing with pipe failures was translated into financial terms by multiplying the estimated time spent by the survey respondent by the estimated hourly wage rate for each zip code. The estimated hourly wage rate was calculated by employing annual average income for each zip code as reported by Census Bureau divided by the assumed number of hours worked (52 weeks*40 hours) in a year (Census Bureau, 2000). The estimated total expenses (financial and time costs) ranged from $1,271 to $18,455. Fifty-six observations had missing values for either the time or money spent on pinhole leak repairs. As a result, the total number of unweighted observations for total costs is 392. Total expenses were regressed on number of leaks, type of pipe material (copper vs. others; iron vs. others), type of water conveyed (hot vs. cold), place of leak in the dwelling, age of the dwelling, and source of water (Patuxent water treatment plant vs. other).

Unstandardized and standardized coefficients of each variable were reported for the weighted ordinary least squares regression. Unstandardized coefficients represent the actual effect of the independent variable on the dependent variable. Standardized coefficients explain the impact of a one standard deviation increase in the independent variable on the dependent variable relative to other variables in the model. As a result, factor standardization allows for observing the relative importance of explanatory factors, while keeping the variable of interest in its original form (Kim and Feree, 1981).

Sample Selection Bias. Information on the total repair costs associated with pinhole leaks might not be available for all the survey respondents and the distribution of respondents over categories of the independent variables may not be random. According to the prediction accuracy analysis (Table 6) based on the logistic regression of pinhole leak occurrences (Table 5), 40 percent of the unweighted survey observations for pinhole leaks incidents were incorrectly classified. The discrepancy between observed and predicted values could be related to differences in unmeasured and measured characteristics between groups with and without pinhole leaks. The unobserved variables might be important in explaining pinhole leak occurrences (Smits, 2003). The unobserved factors contributing to misclassification of observed data might have also biased the estimates of least squares regression for the total expenses to repair pinhole leak damage. A Heckman two-stage procedure was utilized to correct for the selection bias in order to estimate the unbiased effects of independent variables on total pinhole corrosion repair expenses (Briggs, 2004).

Under the Heckman model, the following ordinary least squares regression assumptions are retained: i) the error term (e_i) is independently and identically distributed in i with a standard normal distribution; and ii) $\{x_i: i, \ldots, n\}$ are independent of $\{ e_i: i = 1, \ldots , n\}$,. Additionally, two further statistical assumptions are made: iii) δ_i is independently and identically distributed in i with a standard normal distribution; and iv) $\{\mathrm{x}_i: i = 1, \ldots , n\}$ is independent of $\{\delta_i: i = 1, \ldots , n\}$. The OLS assumption of independence between error terms (e_i) and unmeasured latent characteristics (δ_i) in respondent i is violated (Briggs, 2004).

Allowing for the correlation between error terms from the total costs regression and unmeasured characteristics of each respondent ($Corr(e, \delta) = \rho$) introduces a correlation variable (ρ) to the model specification. When correlation factor is zero, then there is no selection bias and estimation results are unbiased. On the other hand, if unobserved characteristics impact pinhole leak occurrences, then the correlation coefficient is different from zero (takes values between –1 and 1) and the least squares regression results are biased. Additionally, it is impossible to estimate the OLS regression without considering the selection process. As a result, the Heckman model utilizes a two-stage process involving a "selection equation" and a "substantial equation" in which both equations are related to each other to account for the unobserved characteristics. The correlation between the error terms of selection and substantial equation should be estimated to confirm the sample selection bias occurrence in the model specification (Briggs, 2004; Smits, 2003).

In the first step of the Heckman procedure, the selection equation was estimated to account for selection bias. Bias of least squares coefficients resulted from not accounting for the differences in unmeasured characteristics between people reporting pinhole leaks and those not reporting. The weighted binary logistic model used to predict leak occurrences as described in equations (1) and (2) was employed to compare these dissimilarities between the two groups. Because the Heckman model assumes a normally distributed error term and the logistic regression does not, the logistic model estimates were converted into probit model estimates. Using the inverse cumulative distribution function of the normal distribution, logistic individual probabilities (ranging from 0.17 to 0.81) were translated into the individual probit scores (ξi) for each respondent. The probit scores represent the residuals between the observed and predicted probabilities and range between -0.94 and 0.86. The residuals of the selection equation were then utilized to construct a selection bias correction, which is equivalent to the inverse Mill's ratio. The following formulas were employed to calculate the sample selection bias correction factor:

$$\lambda_i = ((1/sqrt(2*3.142))*(exp(-\xi_i * \xi_i *0.5)))/ \Phi_i (\xi_i), \text{ for } y = 1 \qquad (5)$$

$$\lambda_i = - ((1/sqrt(2*3.142))*(exp(-\xi_i * \xi_i *0.5)))/(1- \Phi_i (\xi_i)), \text{ for } y = 0 \qquad (6)$$

where ξ_i are the individual probit scores. The selection bias correction measures the effects of all unmeasured characteristics of respondents related to pinhole leak corrosion. The bias correction variable is added as an explanatory variable. As a result, the remaining variation in the dependent variable after the removal of the effect of the known variables is explained by the unknown factors (Smits, 2003).

The inverse Mill's ratio for each respondent is explained as follows:

$$\lambda_i = \phi_i(Z) / \Phi_i(Z), \text{ for } y = 1 \quad (7)$$

$$\lambda_i = -\phi_i(Z) /(1- \Phi_i(Z)), \text{ for } y = 0 \quad (8)$$

where Z is the vector of explanatory variables for the respondent, respectively; ϕ_1 is the density probability function; and Φ_1 is the cumulative probability function. If the predicted probability of pinhole leak occurrences is close (but not equal) to unity for respondents with pinhole leaks, this implies that $\Phi_i(Z) \rightarrow 1$ and that $\phi_i(Z) \rightarrow 0$. As a result, the inverse Mills ratio term is close to zero for this respondent and the selection bias correction variable has no influence on the total cost of pinhole leak damage. In case of respondents without pinhole leaks occurrences, the high-predicted probability of corrosion implies high selection bias in absolute value. As a result, these respondents should have experienced pinhole leaks in their plumbing system. On the other hand, if the predicted probability is small for respondents with pinhole leaks occurrences, the inverse Mills ratio is large ($\Phi_i(Z) \rightarrow 0$ and $\phi_i(Z) \rightarrow 1$) and the selection bias correction variable strongly influences the total expenses of corrosion damage. For the respondents without pinhole leaks, the small-predicted probability implies the inverse Mill’s ratio to be small in absolute terms. The low bias in absolute value means that respondents without pinhole leaks appropriately did not observe pinhole leaks in their plumbing (Briggs, 2004). For this analysis, the selection bias variable was calculated for 879 unweighted observations and ranged between -1.42 and 1.44.

In the second step of the Heckman procedure (substantial analysis), the weighted ordinary least square regression for total repair expenses was performed. The selection bias correction factor was added as an additional explanatory variable of pinhole leak repair costs. The ordinary least square model specification with Heckman selection bias correction is provided by

$$y = X\boldsymbol{\beta} + \lambda\boldsymbol{\theta} + \boldsymbol{e} \qquad (9)$$

where X is the vector of explanatory variables for the respondent with pinhole leaks, λ is the sample selection bias correction for respondents with pinhole leaks, and θ is the coefficient of the selection bias defined as $\theta = \rho * \sigma$.[5] Because the selection bias correction variable reflects all of the unmeasured characteristics related to pinhole leak occurrences, the remaining independent variables are free from this effect and unbiased. On the other hand, the standard errors of the ordinary least square estimates are heteroskedastic:

$$\boldsymbol{e} \sim (0, \sigma^2 V_i)) \qquad (10).$$

With heteroskedastic errors, the least square estimates are still unbiased, but not efficient. As a result, the variance varies with each respondent (the true variance is either overestimated or underestimated). In order to correct for heteroskedasticity (obtain constant variance: $\boldsymbol{e} \sim (0, \sigma^2 I_i))$, a generalized least square regression (GLS) was employed (Smits, 2003).

In order to confirm the existence of selection bias in the total costs model specification, the correlation between the error terms of the selection and substantial equations should be estimated ($Corr(e, \delta) = \rho$). As presented in equation 9, the coefficient of selection bias estimated in the substantial analysis is defined in part by ρ. As a result, the correlation coefficient is calculated as follows:

$$\rho \;=\; sqrt(\hat{\theta}^2 / \hat{\sigma}^2), \qquad (11)$$

[5] To test the significance of the selection bias correction factor in the model specification, the F-test was employed. The F-test examines the importance of correction for selection bias by comparing models with and without sample selection bias correction variable. The F-test statistic is defined as follows:

$$F\text{-}test\ statistic = (SSE_r - SSE_u) / (SSE_u /(n\text{-}k)) \sim F\ (1, (n\text{-}k)), \qquad (9a)$$

where SSE_r is the error sum of squares in the model without sample selection bias, SSE_u is the error sum of squares in the model with sample selection bias, n is the number of observations in the model with selection bias, and k is the number of coefficients in the model with sample selection bias. The hypotheses are as follows:

H_0: $y = X\boldsymbol{\beta} + \boldsymbol{e}$

$H_{1:}$ $y = X\boldsymbol{\beta} + \lambda\boldsymbol{\theta} + \boldsymbol{e}$.

If at a 5% significance level and (1, n-k) degrees of freedom, the critical F-value is less than the F-test statistic, then the H_0 is rejected and the sample selection bias correction should be included in the model (Griffiths et al., 1993).

where $\hat{\theta}$ is the estimated coefficient of selection bias correction and $\hat{\sigma}^2$ is the estimated variance from the substantial analysis (Smits, 2003).

Results without accounting for selection bias correction. The weighted ordinary least squares estimation was performed on the observations that reported pinhole leaks. The total unweighted number of observations utilized to estimate the effect of independent variables on the total costs is 392 after deleting 56 observations with missing data for total expenses of pinhole leak damage. According to the OLS results, the number of leaks had a large positive impact on pinhole leak costs (Table 9). Copper plumbing and pipe failure on the first and second floor of a dwelling increased repair expenses. Plumbing system malfunctions under the slab and in the basement, on the other hand, were negatively related to the total repair expenses, probably because these areas were easier to access for repair and/or because leaks in these areas represented less associated damage to the structure and furnishings of the dwelling. Iron pipes, cold water pipes, dwelling age, and source of water were not related to pinhole leak damage costs.

Results accounting for selection bias correction. The effective unweighted sample size after employing the selection bias correction in the least squares regression is 367. The sample selection bias correction values are missing for 25 observations. The missing observations relate to the respondents' information not included in the selection equation due to incomplete data of explanatory variables. As a result, these observations were not integrated into the logistic regression estimation and have no selection bias correction value. The selection bias for respondents with corrosion ranges between 0.38 and 1.44. The correlation coefficient, which represents the error term correlation between the selection and substantial equation, is 0.44 with standard deviation equal to 0.00. The F-test statistic, which tests the importance of selection bias correction in the model specification, is 7.86 (critical $F_{(1, 356)}$ value = 3.84). The above finding confirms the need for sample selection bias correction in the total expenses of corrosion damage model specification.

According to the results (Table 9), the selection bias correction coefficient is 0.11, but it is not significant at the 0.05 significance level. However, including the selection bias caused coefficients and significance levels of other variables to change. Compared to the regression without selection bias correction, the magnitudes of the standardized coefficients increased in absolute value for the following variables: copper pipes (from 0.06 to 0.13) and leak under slab or in basement (from -0.19 to -0.20). The magnitudes of the standardized coefficients decreased in absolute value for number of leaks (from 0.26 to 0.24), leak on the first floor (from 0.29 to 0.25), and leak on the second floor (from 0.19 to 0.18).

Similarly to the ordinary least squares model without selection bias correction, number of leaks, copper plumbing and leaks on the first and second floor of a dwelling increased repair expenses while leaks under the slab and in the basement are negatively related to total repair expenses. Additionally, water from the Patuxent treatment plant is negatively related to the repair costs of pinhole leaks. One possible explanation for this result, among many, is differences in water treatment procedure between competing water treatment plants in the area (Patuxent treatment plant water might be less corrosive). Dwelling age, cold water pipes, and iron pipes are still not significant in the corrected model.

Table 9: Total Costs of Repairing Pinhole Leaks and Associated Damage (Least Squares Regression)

	OLS Regression without accounting for selection bias correction			GLS Regression with accounting for selection bias correction		
Variables	Unstandardized Coefficients	Standardized Coefficients	t-Statistic	Unstandardized Coefficients	Standardized Coefficients	t-Statistic
Constant	-961.76		-1.6	-2892.86		-0.1
Number of leaks	785.49	0.26	5.1[b]	688.34	0.24	5.3[b]
Copper pipes	664.24	0.06	1.7[c]	1287.27	0.13	1.8[c]
Iron pipes	237.21	0.02	0.3	513.55	0.04	-0.8
Cold water pipes	-258.54	-0.02	-0.4	3.92	0.00	0.5
Leak under slab and in Basement	-1462.04	-0.19	-3.0[b]	-1441.72	-0.20	-2.1[c]
Leak on the first floor	2141.77	0.29	5.2[b]	1740.57	0.25	5.9[b]
Leak on the second floor	1697.72	0.19	3.9[b]	1512.92	0.18	4.1[b]
House built before 1960	-315.03	-0.04	-0.9	-143.03	-0.02	-0.6
Source of water	-671.00	-0.06	-1.3	-85.40	-0.01	-3.6[b]
Selection bias correction	N/A	N/A	N/A	1597.99	0.11	0.1

[a] Observations weighted by sample weights.

[b] Significant at 5%.

[c] Significant at 10%.

[d] Regression without accounting for selection bias correction: $R^2 = 0.231$; F-stat = 12.48; n (unweighted) = 392.

[e] Regression with accounting for selection bias correction: $R^2 = 0.281$; F-stat = 13.93; Effective unweighted sample size after selection bias correction n = 367.

Sensitivity Analysis of Housing Age

Potentially results of the regression analyses might be influenced by the large number of dwellings (498) built in the 1950-1969 period. An additional weighting was done to reduce the influence of the 1950-1969 period on regression results. The 1950-1969 period was made the base period and respondents living in homes built then received the weight based on their initial sampling weight. Respondents living in homes from other periods were weighted in inverse proportion to the ratio of homes in their age category to the base number of homes. For example, there were 26 respondents living in dwellings built since year 2000. Therefore, the weight for respondents living in dwellings built since 2000 is 498/26 = 19.15. This weight was multiplied by the initial sampling weight described above. For example, a respondent living in a zip code receiving a weight of 1.5 and in a dwelling built after 2000 would receive a final weight of 19.15 times 1.5 = 28.7. The unweighted sample size without selection bias correction is 392 (OLS regression) and with selection bias correction is 367 observations (GLS regression).

The results of the logistic model for observations weighted by dwelling age category (Table 5) are different from the base model. Probabilities associated with water travel time, history of pipe failure, pipes replaced between 1960 and 1980, source of water, and use of copper pipes do not differ much from the base model. However, the variable for a dwelling built before 1970 is significant at the 0.05 level in the housing age-weighted model while not significant in the base model. The factor's magnitude is also lower in the housing age-weighted model compared to the base model (50 percent vs. 52 percent). Consequently, weighting for the number of dwellings built between 1950 and 1969 picks up the impact of dwelling age on pinhole leak occurrences.

In the case of ordinary least squares regression of total expenses with no correction for sample selection bias (Table 9 and 10), the results are very similar between the base model and the model weighted by the age category of dwelling. Similarly to the base model, number of leaks, copper plumbing and leaks on the first and second floor of a dwelling increased repair expenses while leaks under the slab and in the basement are negatively related to total repair expenses. Dwelling age, water source, cold water pipes, and iron pipes are still not significant. Compared to the base model, the magnitude of the standardized coefficient decreased in absolute value for the number of leaks variable (from 0.26 to 0.25).

There are, however, differences between the least squares estimates without selection bias correction and the results accounting for the bias (Table 10). The selection bias correction coefficient is 0.02 and statistically insignificant while the correlation coefficient between the error terms of selection and substantial analysis is 0.15 with standard deviation equal to 0.00. The F-test statistic, which tests the importance of selection bias correction in the model specification, is 8.10. The above finding confirms the importance of selection bias correction inclusion in the model specification. Similarly to the base model, water from the Patuxent treatment plant became statistically negatively related to the repair costs of pinhole leaks after accounting for the selection bias. Compared to the regression without selection bias correction, the magnitudes of the standardized coefficients increased in absolute value for the following variables: number of leaks (from 0.25 to 0.27), leak on the first floor (from 0.29 to 0.31) and leak on the second floor (from 0.19 to 0.21). The magnitude of the standardized coefficient decreased in absolute value for leaks under slab or in basement (from -0.19 to -0.11).

Table 10: Total Costs of Repairing Pinhole Leaks and Associated Damage in the Age-Weighted Model (Least Squares Regression)

	Observations weighted by age category of dwelling					
	OLS Regression without accounting for selection bias correction			GLS Regression with accounting for selection bias correction		
Variables	Unstandardized Coefficients	Standardized Coefficients	t-Statistic	Unstandardized Coefficients	Standardized Coefficients	t-Statistic
Constant	-920.67		-1.6	-667.79		-0.3
Number of leaks	766.09	0.25	5.1[b]	774.48	0.27	5.4[b]
Copper pipes	605.68	0.06	1.7[c]	798.13	0.06	1.8[c]
Iron pipes	333.04	0.02	0.5	-461.43	-0.03	-0.7
Cold water pipes	-295.47	-0.02	-0.4	-302.99	-0.02	-0.4
Leak under slab and in Basement	-1537.36	-0.19	-3.1[b]	-996.21	-0.11	-2.1[c]
Leak on the first floor	2174.85	0.29	5.3[b]	2459.96	0.31	6.0[c]
Leak on the second floor	1727.72	0.19	4.0[b]	1997.96	0.21	4.4[b]
House built before 1960	-3.37	-0.06	-1.4	-0.49	-0.01	-0.1
Source of water	-778.80	-0.07	-1.5	-1844.85	-0.20	-3.1[b]
Selection bias correction	N/A	N/A	N/A	466.43	0.02	0.2

[a] Observations weighted by sample weights and age of dwelling as described in text.

[b] Significant at 5%.

[c] Significant at 10%.

[d] Regression without accounting for selection bias correction: R^2 = 0.234; F-stat = 12.83; n (unweigthed) = 392.

[e] Regression with accounting for selection bias correction: R^2 = 0.288; F-stat = 14.49; Effective unweighted sample size after selection bias correction n = 367.

Conclusions

A survey investigating the pinhole leaks problem in the Maryland suburbs of the Washington, D.C. area was conducted in 2004. Results from the survey indicate that pinhole leaks in home plumbing systems are a frequently reported problem in this area. This region was predicted to have unusually a high pinhole leak activity at the start of the project. Most pinhole leaks occurred in the mid-1990's and the 2000's. Most respondents had 1 to 4 leaks; however, 8 percent (weighted basis) reported 7 or more leaks. Probability of pinhole leaks increases with use of copper pipes as the plumbing material and as the dwelling is located closer to the treatment plant. Leaks are also more likely to occur if a plumbing system has failed in the past. Pinhole leaks can have high financial, emotional, and time costs. Repair costs increase with the number of leaks, repairs of copper pipes, and location of leaks on the first or second floor of the dwelling. Accounting for sample selection bias affected the level and statistical significance of variables related to cost of repairing pinhole leak damages.

Due to the increasing frequency of pinhole leak occurrences across the nation, the Virginia Tech researchers working on this NSF-funded project have extended the investigation of pinhole leaks nationwide. Further research is being done on factors associated with pinhole leak occurrences in home plumbing and costs of pinhole leak damage. Households' preferences for plumbing materials and their willingness to pay for improved plumbing infrastructure are also being examined. The research findings will inform policymakers, program managers, and water utilities on the importance of reducing corrosion in home drinking water infrastructure.

References

D. C. Briggs. 2004. “Causal Inference and the Heckman Model.” *Journal of Educational and Behavioral Statistics* 29(4): 397-420.

Census Bureau. 2000. “Quick Facts.” Available online at www.census.gov. Accessed March 2006.

M. Edwards, J.C. Rushing, S. Kvech, and S. Reiber. 2004. “Assessing Copper Pinhole Leaks in Residential Plumbing.” *Water Science & Technology* 49(2): 83–90.

EPA. 1992. “Drinking Water Standards for Regulated Contaminants.” Available online at http://www.epa.gov/safewater/therul.html. Accessed January 2004.

S. Fleishman. 2001. “And the Leaks Go On.” *Washington Post*: G01.

D. G. Garson. 2006. “Logistic Regression.” Available online at http://www.2.chass.ncsu.edu/garson/PA765/logistic.htm. Accessed April 2006.

W. E. Griffiths, R. C. Hill, G. G. Judge. 1993. *Learning and Practicing Econometrics*. John Wiley & Sons, Inc.: New York.

S. Gurner. 2003. “Pinhole Leaks In Copper Pipes.” Available online at http://strathmore-belpre.org.

J. Kim and G. Feree. 1981. “Standardization in Causal Analysis.” *Sociological Methods Research* 10(2): 187-210.

E. J. Kleczyk, D. J. Bosch. 2006. “Corrosion in Home Plumbing Systems: Assessment of Costs and Causal Factor.” The 2006 AREUEA Mid-Year Meeting. Washington, D.C.

E. J. Kleczyk, D. J. Bosch, S. Dwyer, J. Lee, and G.V. Loganathan. 2005. “Maryland Home Drinking Water Assessment.” The Proceedings of the 2005 Virginia Water Research Symposium. Virginia Water Resources Research Center. Virginia Tech, Blacksburg, Virginia. 104-113. Available online at www.vwrrc.vt.edu.

J.C. Rushing and M. Edwards. 2004. “Effect of Aluminum Solids and Free Cl_2 on Copper Pitting.” *Corrosion Science* 46(12): 3069-3088.

J. Smits. 2003. “Estimating the Heckman Two-Step Procedure to Control for Selection Bias with SPSS.” Available online at http://home.planet.nl/~smits.jeroen. Accessed April 2006.

UCLA Academic Technology Services. 2005. “Annotated SPSS Output Logistic Regression.” Available online at www.ats.ucla.edu/STAT/SPSS/output/logistic.htm. Accessed April 2006.

Washington Suburban Sanitary Commission (WSSC). No date given. “Pinhole Leaks in Copper Pipes.” Available online at http://www.wssc.dst.md.us/copperpipe/pinhole_charts.cfm. Accessed April 2006.