Doing Multivariate Analysis of Residential Single-Family Water Conservation Programs

THE COOKBOOK

Morrison Institute for Public Policy

School of Public Affairs

Arizona State University

Doing
Multivariate
Analysis of
Residential
Single-Family
Water
Conservation
Programs

THE COOKBOOK

Heather E. Campbell, PhD
School of Public Affairs

■

Ryan M. Johnson, MPA
Morrison Institute for Public Policy

■

ASU
ARIZONA STATE UNIVERSITY

MORRISON INSTITUTE FOR PUBLIC POLICY
School of Public Affairs, College of Public Programs, Arizona State University
PO Box 874405, Tempe, AZ 85287-4405 ■ voice 480.965.4525 ■ fax 480.965.9219 ■ http://www.asu.edu/copp/morrison

SCHOOL OF PUBLIC AFFAIRS
College of Public Programs, Arizona State University
PO Box 870603, Tempe, AZ 85287-0603 ■ voice 480.965.3926 ■ fax 480.965.9248 ■ http://www.asu.edu/copp/publicaffairs

Design by Chalk Graphic Design ■ chalk@primenet.com

CONTENTS

The Morrison Institute for Public Policy is an Arizona State University resource for objective public policy analysis and expertise. The Morrison Institute researches public policy issues, informs policy makers and residents and advises leaders on choices and actions. The Morrison Institute is a unit within the School of Public Affairs, one of the departments of the College of Public Programs.

The School of Public Affairs is a NASPAA-accredited comprehensive school offering degrees in Public Administration. The mission of the school is to serve the public policymaking needs of Arizona and the nation, prepare students for leadership in public service, and advance research in discovering, integrating, sharing, and applying knowledge of public affairs. The School performs these activities through its teaching programs, the research of its faculty, the research of the Morrison Institute, and the training activities of the Advanced Public Executive Program.

This work was funded by a grant from the Arizona Department of Water Resources.

Introduction

You don't have to be a great chef to put together a good meal.
You don't have to invent a recipe to cook dinner.

Similarly, we believe you shouldn't have to know how to develop
a statistical model to produce a good water evaluation.

This book is designed to be a "cookbook" for water evaluators who would like to be able to do a good job evaluating their water programs for decision-making, but who don't know a lot about statistics.

For most people, statistics is like a foreign language. So, we've used the extended metaphor of a cookbook about a foreign cuisine. With a foreign food cookbook, not only are there recipes, but there is a discussion of the staples of the cuisine—what you need in your larder or pantry. Since sometimes, especially if you don't live in a large city, you won't have access to all the exotic ingredients, substitutions are given. You don't have to understand all the ingredients or precisely their function in the recipe (cream of tartar, anyone?), but if you follow the recipe diligently, you should end up with something satisfying and useful.

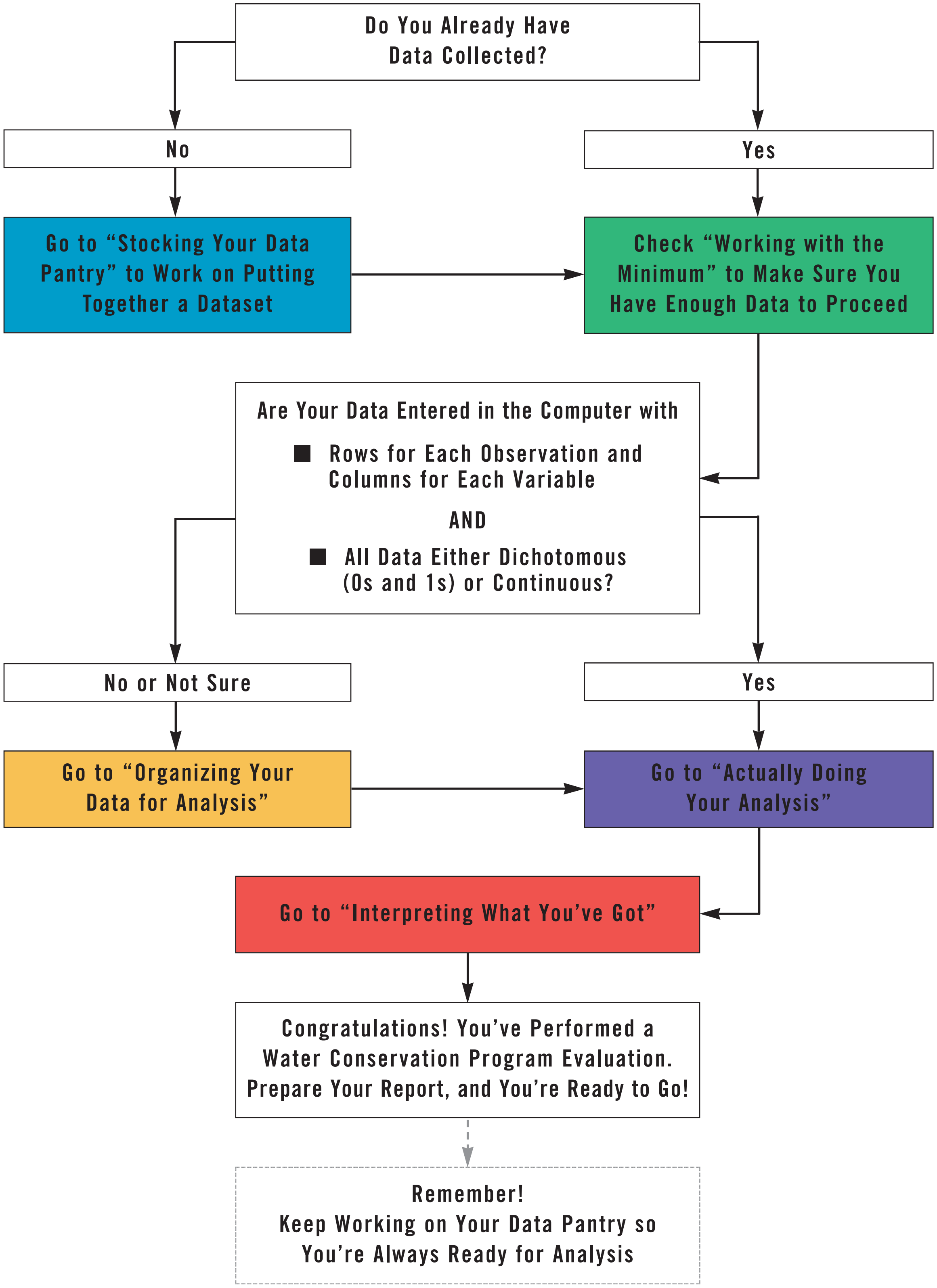
As with a cookbook for a foreign cuisine, we'll tell you what you need in your "data pantry" so that you can do the best job of your analysis—even when you have unexpected demands. We'll also talk about substitutions—what to do when the best ingredients aren't at hand, but you still need to present something to a hungry audience. If you don't know much about statistics, making up your own substitutions may decrease the value of the outcome, but if you follow the recipes provided here, you should have something useable.

You should also keep in mind that, like any cookbook, this one doesn't include all the recipes possible for developing analyses of water conservation programs. This one focuses on multivariate regression analysis, using a model developed by the authors for the Arizona Department of Water Resources and the City of Phoenix. Multivariate regression analysis is by no means the simplest technique you could use, so why is that what we're starting with? Because it is the best technique for figuring out the independent effects of your water conservation programs. Simpler techniques are much less powerful in detecting what you really care about, and can sometimes even be deceptive. But don't worry: the multivariate regression this cookbook guides you through is fairly easy as these things go—it is not the gourmet version, but it is effective at teasing out the independent effects of factors that cause water consumption.

Someone who actually understands the statistical techniques used here would be better able to evaluate water programs. Maybe using this cookbook will make you want to go learn more about certain techniques, and we would be delighted with that; but mostly, we hope that you will use this and find it useful.

When guests come into our homes, they can easily tell which cookbooks are our favorites by which are the dirtiest—they're the ones that have been splashed on, spilled on, and used over and over. We hope that you'll give *Doing Multivariate Analysis of Residential, Single-Family Water Conservation Programs: The Cookbook* such a workout, and that it will become a useful tool for you in assuaging hunger about the value of different water conservation programs. Dog-ear it, underline it, write on it—but whatever you do, please USE IT.

Using This Book


Do You Already Have Data Collected?
No
Yes
Go to “Stocking Your Data Pantry” to Work on Putting Together a Dataset
Check “Working with the Minimum” to Make Sure You Have Enough Data to Proceed
Are Your Data Entered in the Computer with
■ Rows for Each Observation and Columns for Each Variable
AND
■ All Data Either Dichotomous (0s and 1s) or Continuous?
No or Not Sure
Yes
Go to “Organizing Your Data for Analysis”
Go to “Actually Doing Your Analysis”
Go to “Interpreting What You’ve Got”
Congratulations! You’ve Performed a Water Conservation Program Evaluation. Prepare Your Report, and You’re Ready to Go!
Remember! Keep Working on Your Data Pantry so You’re Always Ready for Analysis

Notice there’s a glossary, if you don’t know some of these terms.

Stocking Your Data Pantry

This section discusses the following topics, arranged alphabetically:

Computer and Software

You can't cook without a stove or oven, and you can't perform modern statistical analysis without a computer and software. In order to perform a statistical analysis of water conservation measures, you should have a computer equipped with the following:

1. Software, such as Microsoft® Excel, that allows you to enter and store data in matched rows and columns with descriptive names.
   - You can get by without this, but it makes the task harder. Data still have to be entered in rows and columns, but you can do this "typewriter style" with tabs between columns.
2. Software, such as SPSS®, that allows you to perform statistical analysis of data.
   - You can't get by without a statistical package, but can use many other packages such as TSP® (powerful and inexpensive, but not slick), SAS® (powerful but cumbersome), STATA® (powerful and with good teaching texts), or even Minitab® (very old and simplistic) in a pinch. In fact, the authors found SPSS to be somewhat limiting, but use it as an example because many local governments already use it.
3. Reasonable capacity for storing your data.
   - Depending on the type of analysis you want to do, "reasonable storage capacity" can vary from 10 megabytes to 100 megabytes. Right now, above 100 megabytes is beyond "reasonable," but can be useful. (In 5 years or less, 100 megabytes will probably seem laughably little.)

Please note that you do not have to use some huge intimidating mainframe that is controlled by the priests of information technology (IT). Given current computer power, you can perform even the most complex analysis discussed here on a modern laptop computer—though you may have to begin by requesting information from your IT people. For under $5000 (in 1999), you can have the capacity to store and analyze all the data you are likely to want to deal with, and the freedom to add data and perform analysis anytime you wish.

Data Storage in 1999

It used to be that data storage was a big problem, especially for small jurisdictions. This should no longer be the case. Even a small district should be able to purchase an inexpensive computer just for the purpose of storing data to be used for water conservation program evaluation. Since even inexpensive machines come with more than a gigabyte (1,000 megabytes) of storage capacity now, this will store a great deal of data for analysis—along with the software for entering and analyzing the data. If you don't have other data storage capacity, buy a laptop, and think of it as your water program evaluation file cabinet—your data pantry. Furthermore, if you begin to run out of space, Zip® drives are relatively inexpensive (around $100) and each Zip disk holds 100 megabytes. Jaz® drives are somewhat more (around $200 to $400), but each Jaz disk stores 1 gigabyte. They are easy to connect to your computer and even easier to use.

Detail

A good data pantry contains a great deal of detailed information. It is always a good principle of data collection to collect more data than you think you will need, to store data at the most disaggregated level possible, and to store data for as long as possible (try for at least 5-7 years). Three mistakes are very common when organizations want to evaluate programs. The first is that they haven't collected the data that allow them to measure what they care about. The second is that they have stored data at an aggregated level and thrown away the rest. The third is that they have collected data, but not over a long enough time period.

Good Principles of Data Collection

1. Collect more data—both numbers and contextual narrative—than you think you will need, and do it as soon as possible.
2. Collect data that are designed to help you answer questions you care about.
3. Store data at the most disaggregated (smallest) level possible. Example: don't store averages when you can store the individual elements you would use to make the averages.
4. Store data for long as possible (try for at least 5-7 years).

Please understand that the word "data" really means "information." Don't just store numbers; also store a narrative about the program, what was done, and how it worked. You may think you or the others involved will remember later, but it often won't work that way. Work goes on; people forget, and sometimes people leave and you can't get hold of them. So, not only is it important to do this, but it is important to do it SOON while memories are still fresh.

- *A comment on how much data you’ll store:* Please don’t think that we are suggesting that you will store all your billing records here. You don’t need the whole cow to make a steak.
    - The most rigorous standard for statistical estimation requires about 1200 randomly selected observations. For example, for one year, you could have data on 1200 customers, some of whom received a particular program and some of whom didn’t. For these 1200 people, you want all 12 months of their usage information. Or, if you wanted to go all out, you could have 1200 who did receive a program and 1200 who didn’t (again, keeping all 12 months). More—other than to cover more years—just isn’t necessary.
    - And, there are also much less rigorous standards. For example, if you have about 33 more observations than variables in your model, you can still have reasonable confidence in your results.
    - Whatever you do, don’t let the ideal be the enemy of the doable! Do you store every spice in the world? No, only those you use often. **Don’t try to store so much data that you’re exhausted just thinking about it, or too intimidated to work with it.**
    - BUT if you work for a water provider that destroys billing records after 18 months, get a random sample of records every period before they’re destroyed, and store them in your data pantry. Then, you’ll be ready if an unexpected need for analysis comes up. One way to do this is to randomly select 1200 accounts and keep pulling those same folks every storage cycle. But remember, if your community has been growing rapidly since you started to do this, you’ll need to add some random selections from the newer accounts.

DON’T Be Overwhelmed; DO Be Creative

The preceding discussion of 1200 accounts, and the next discussion of 12 months of data might seem overwhelming. Maybe you feel you just can’t meet those goals. *Don’t let that stop you! Don’t let the ideal be the enemy of the doable.*

<u>Look for the minimum requirements.</u> If you can’t do 1200, can you do 100? How about 48? 48 accounts meets the “33 more observations than variables” rule with a 15-variable model. More is better, but this is respectable. If you can’t do 12 months, how about 4, 2 before some accounts have received a program and 2 after? Keep thinking! Even if based on less data than the ideal, actual estimates can really help out when what you’ve had is a hunch or a hope.

<u>Be creative.</u>

- If your billing division is overwhelmed with requests from many different groups, maybe you can piggy-back on another group’s data request. It might be a lot easier to pull all the utility info for 1200 accounts than to pull water for 1200 and electricity for 1200, and....
- Maybe you can do a small survey. Could you afford to survey 100 households twice? You could survey 100, then give 50 a program, and then survey the 100 again (when designing the survey, focus on collecting information on elements you want in your data pantry). If you ask how much water was on the bill the month before, you don’t have to get anything from billing.
- If you can’t imagine finding the time to enter data, consider a student intern. Students will often work for little money or course credit, because they want experience.
- Consider pooling data with other cities. Suppose Mesa has data for 100 accounts, and so do Tempe and Chandler. Now you’ve got a 300-account pool—and maybe more programs (but do add controls for important differences *between* providers, if any).

Keep thinking, and keep storing data. Maybe you don’t have enough now, but if you keep trying, maybe you’ll have enough eventually.

Information to Store

For Your Community

You should store total water use, preferably broken down by sector (that is, industrial use separated from residential and agricultural, etc.). Store this as frequently as you can—every year at least, quarterly or monthly if possible. You should also store population data and/or total number of accounts—preferably at the same frequency as the total water use data; try for at least yearly.

For All Accounts to Be Used in Analysis

Try to store 12 months of water consumption data for every account that you would like to analyze. (As discussed below under "Variability," this includes for some customers who haven't yet received any water conservation programs.) Also store information, such as address, that lets you link to US Census tracts and to locate the closest weather stations.

For All Conservation Programs

Ideally, for every water conservation program you implement, whether delivered to individuals or across all customers, you should retain records of the amount spent. This allows you to go beyond evaluating whether programs work or not, to evaluating how much water savings you get per dollar spent. But do note that this is the ideal. Just because you don't have cost records doesn't mean you can't perform a useful evaluation.

For Individual-Level Conservation Programs

For every water conservation program that you implement at an individual level—that is, that is delivered to individual people or households rather than to everyone in your district—you should always collect and store information that will allow you to link the people who received the service to their water account. For example, always make sure to get the addresses of people who participate in individual-level water conservation programs.

- Individual-level programs might include Xeriscaping® (desert landscaping) rebates, Xeriscaping seminars, toilet rebates, fixture deliveries. They are programs that are delivered at some level less than to everyone in the water provider's district. Note that this may include programs that are *offered* to everyone in the district, but the key here is that you can tell which individuals received them. Xeriscaping rebates may be an example if they are available for everyone in the community to take advantage of, but not everyone does, and you can tell who does.
- For some programs, this may be a bit confusing. What about teacher education programs? For these programs, it's more important to get the addresses of the kids the teacher teaches than of the teacher her/himself. But, to get that information later, you need to know how to contact the teacher!

For Provider-Wide Conservation Programs

For every water conservation program that you implement for all customers, you should keep careful track of the start and end dates for the program.

- Programs targeting "all customers" might include radio or TV education campaigns, making evapotranspiration (ET) data readily available to customers, or a new rate-structure. They are programs that are delivered to everyone, or that are delivered so broadly that you can't tell who received them and who didn't.

About Other Things

When it comes right down to it, the intent of water conservation programs is to reduce water consumption. But, as you know very well, many things effect water consumption other than your conservation programs. In order to be able to tell what effect your water conservation program has had, you need to be able to take into account the effect of those "Other Things." Here are the most common other things that affect water conservation and that you should keep records of.

- Climate Data: It is well known that heat and rainfall (precipitation) affect water use. For each time period that you have consumption data stored, try to store data on evapotranspiration (ET) and precipitation for your community. These data should be easy to come by. If you don't have another source, AZMET, the Arizona Meteorological Network, provides data and links to other data sources at *Ag.Arizona.Edu/AZMET*. If your community is big enough to have important variation in ET and precipitation at the SAME time period BETWEEN parts of the community, match accounts to their closest weather station.

- Pricing Data, in Real/Constant/Base-Year Dollars: Prices affect how much water people use (even if you don't believe this, many others do, and they won't take your analysis seriously if you don't include pricing data). So, for all time periods over which you might want to estimate the effectiveness of water conservation programs, be sure to keep careful records of your water pricing structure.

  Don't forget to include other prices that customers may perceive as water prices, even if you don't think of them that way. For example, is your sewer fee based on water usage during certain months? Then those charges are part of the price of water during those times. Is there an "environmental charge?" That is part of the price, so add it in.

  - You also need to keep or be able to collect data on a price index that controls for inflation. The impact of a price depends on the general level of prices and what inflation has been like. Is a 10-cent increase large or small? Your Mom or Grandma will tell you that when she was little 10 cents would buy a trip to the movies along with the candy! But, in the '70s, when inflation in the US was double-digit, a 10-cent increase was nothing. You have to adjust your prices for inflation and put them in what are known as "Real" or "Constant" or "Base-Year" dollars. You can get price index data many places, including from the Bureau of Labor Statistics website at *www.bls.gov*.

- Household Data: Water usage is measured at the household level, but households may differ in ways that are important to overall water use. The following are factors that are known to be important:

  - Number of members of the household—as the number of people in a household increases, but everything else stays the same, water consumption is expected to increase.
  - Presence of children—when everything else is the same, having children in the house seems to decrease water use compared to a household that doesn't have children.
  - Presence of young adults (those 17-24)—young adults seem to use more water than those in other age groups.
  - Household Income and Poverty—when other things stay the same, it seems that those who are poorer use more water, perhaps because they are less able to afford water-saving fixtures and devices.
  - Whether members of the household are Hispanic, or whether the household is non-English-speaking—in Phoenix, Hispanics seem to use more water than other groups (including other minority groups)[2]. It is not clear if this is cultural or due to language barriers preventing Hispanic households from receiving conservation messages.

[2] For information on this factor and others discussed in this report, see Campbell, Heather E., Elizabeth Hunt Larson, Ryan Johnson, and Mary Jo Waits (January 1999). *Some Best Bets in Residential Water Conservation: Results of Multivariate Regression Analysis, City of Phoenix, 1990-1996*. Morrison Institute for Public Policy, Tempe, Arizona. This research was funded by the Arizona Department of Water Resources, and is available from them or the Morrison Institute.

- Educational attainment—particularly, those with a high school diploma or above use less water.
- Other factors—if you have the resources to collect this information, other household factors that may affect water use are how long people have lived in the community (with longer residents more water conserving), whether people are from the West (with Westerners using less water than those from other regions), whether the household is single-headed (with single-headed households using less water than dual-headed, and female-headed less than male-headed), and whether the home is renter-occupied, or owner-occupied.

  But remember, you don't have to do all of these. Start at the top of the list and measure what you can. Also remember that there may be other factors not listed here that are of particular interest or relevance to your provider or community.
- Where do you get these data?
  - In terms of the quality of the data, the best way is to collect the data from a survey, matched to the accounts in your water conservation program database. However, this way is also expensive and time-consuming.
  - A less precise way that is relatively quick and inexpensive is to use US Census data. Here, you must be able to match Census tracts to accounts (perhaps through address), and your estimates will be less precise because the Census averages must be interpreted as probabilities that the particular account of interest has the average attribute of the tract. This will work better when your community's tracts are fairly homogenous (within specific tracts, households are fairly similar), and less well when they are very heterogenous (within specific tracts, households are very different). Another problem with Census data is that they become out-dated as time passes. If your community is fairly stable, they will work better than if your community has changed a lot since the year of the last Census data available. Census data are available in many different ways, including from the web, at *www.census.gov*.

- House and Yard Data: Houses and yards may differ in ways that are important to overall water use. The following are factors that may matter in water consumption.
  - House age—because newer houses tend to have water-saving fixtures and devices, it is expected that older houses will use more water than newer ones, if everything else is the same. This effect may taper off as older fixtures are replaced.
  - House value—more expensive houses are expected to have more water-using features, including larger landscapes.
  - Number of bathrooms—it is expected that houses with more bathrooms will use more water. (If you can't observe numbers of bathrooms, numbers of bedrooms can be used as a proxy.)
  - Available flood irrigation—houses with available flood (unmetered) irrigation will be *measured* as using less water (because landscape water use will be off-meter).
  - Yard size.
  - Other factors—if there are other house or yard factors you think may be important, such as whether a house has a swimming pool, you can try to collect data on these, as well.
  - Where do you get these data?
    - You may be able to find house age from your account-age information (usually, when a house is built, its water account is started).
    - House value may be available from the assessor's office, and assessor information may also include bathroom information (and information on other features you may think important).
    - Surveys can also gather this information, though people may not want to provide house value information.
    - Sometimes yard-size and yard-feature information can be gathered using aerial maps.
    - If these sources don't work or are too time-consuming, the fall-back is, again, Census data. Census data will not include information on flood irrigation, but there is tract-level information on house age and value, and you can use information on number of bedrooms as a proxy for number of bathrooms. Again, the tract averages must be interpreted as the probability that the account you are analyzing has these characteristics.

Variability

Variability is the most important component of statistical analysis. It is the salt of statistics. You can't cook anything, from brownies to beef, without salt, and you can't do statistical estimation without variability. The more variability in the data, the more information in the data.

Simply stated, variability is difference, or diversity. If you want to get enough nutrients, you need to eat a wide variety of foods. If you want your data to contain a lot of information, you need to get a wide variety of observations.

Variability can come from different sources. The two basic sources of variability are:

1. Observing a single thing (such as a person, a city, or a state) across time; or,
2. Given a single time period, observing "cross-sectionally," that is, across different people or different states or different jurisdictions.

Another powerful technique is to combine cross-time and cross-sectional variation. This is often referred to as "panel" or "pooled" data.

In general, you should always be seeking sources of variability, so, for a single water provider, think in terms of collecting data across time AND across individual water consumers.

Since you need variability in your data, you want to collect data on people who have participated in a water conservation program, and people who have not. You want rich households and poor ones, big households and little ones, old houses and new ones all included in your data. These types of variation are "cross-sectional." If you only have new houses in your dataset, you can't tell how new houses and old ones differ. If you only observe people when they did get the program, you can't tell the effects of the program.

Another way to gain variability is to store data on people (accounts) before they get the program and after they get the program. This is especially useful for programs—such as price structures, or public education messages—that are delivered to all the people in your community. That's why you should start collecting data right away. Just go ahead and take a random sample of your single-family residential water customers. That way, you'll be ready when the time comes.

Q&A: Outliers and Variability

You say that more variability indicates more information, but I've also heard you should throw out "outliers." What's the deal here?

Answer

Variability IS information—as long as the variability is due to real differences. Only throw out "outliers" if you think they're really errors. Example: You know that no house in a community is worth more than $500,000, but one house value is entered as $600,000; you know this is an error; maybe it should have been listed as $60,000, but you're not sure. If you know it's wrong, throw it away. But DON'T just throw it away because it's bigger or smaller than usual. That's throwing away information.

Regarding Variable Ranges

Keep information on the possible range of values. For example, percents cannot be negative; in your community it may be impossible for a single-family residence to use more (or less) than a certain amount of water in one month; etc.

Also, if you use a particular numeric value to indicate missing data, be sure that (1) You keep careful note of this and (2) The value you have chosen is not within the range of values that the variable can meaningfully take on. For example, families can have 9 or 0 children, so don't choose 9 or 0 to indicate missing values for a "number of children" variable.

Figure 1: The IDEAL Data Pantry

1. In 1999, a computer with Excel, SPSS, and 100 megabytes of free space, or the equivalent.
2. For your community:
   - Total residential water consumption for your community, over 10 years or more, at the quarterly or monthly level
   - Population size for your community, yearly (or more often) for 10 years
   - Total number of residential water accounts, at the quarterly or monthly level
   - Climate data—monthly data on ET and precipitation
   - Complete information on all residential water prices
   - Consumer Price Index (CPI) data for putting prices in constant dollars
3. For all water conservation programs:
   - Narrative discussing the program, how it was designed, its purpose, exactly when it started and ended, how much was spent when, exactly how it was implemented
4. For individually delivered water conservation programs:
   - Information that allows you to link recipients to their water accounts
   - Information that allows you to link recipients to climate data
5. Monthly water consumption data for 1200 accounts, some of which have received conservation programs and some of which have not, over 3-7 years.
6. For each of the 1200 accounts:
   - Number of persons in the household
   - Whether there are children in the household
   - Number of people aged 17-24 in the household
   - Household income
   - Whether the household is Hispanic and/or whether English is spoken
   - Educational attainment
   - House age
   - House value
   - Number of bathrooms
   - Size of yard
   - Availability of flood (off-meter) irrigation
   - Other information about houses, property, and people that you think is interesting or important

ORGANIZING YOUR DATA FOR ANALYSIS

This section discusses the following topics:

It's not enough just to have data. You also have to store them in ways that allow you to perform analysis. Ideally, as you are collecting data you will also be entering them into a computer, so that you're always more or less ready to perform analysis with what you've got on hand. This section discusses fundamentals of data storage.

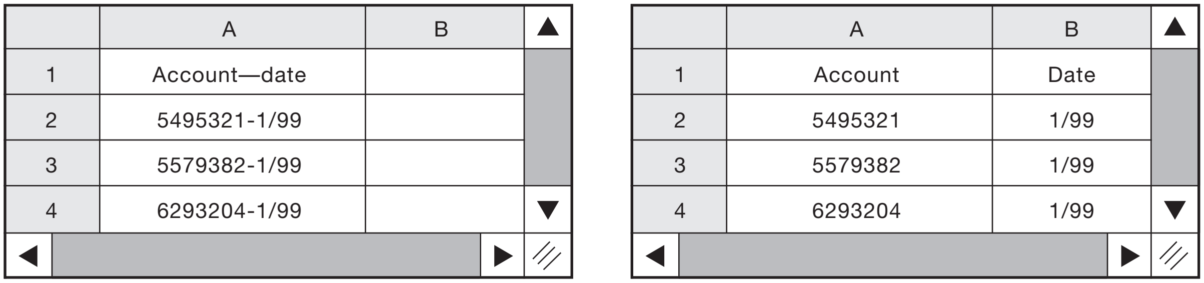
Structuring the Account-Level Data File

To make this easier, let's first imagine a scenario in which you are just starting to collect and store data, so you only have one month's worth of data for several single-family residential customers.

Usually, data should be entered in a computer so that observations are in rows (go across in the data file) and variables are in columns (go down). An "observation" is that unit at which you are observing the data. It's probably easiest to keep focused on your desire to observe water consumption. So, ideally, "an observation" is an account holder. This means that your first column of data should be a list of account numbers. Generally, these account numbers should include some for people who have participated in water conservation programs and some who have not.

It doesn't matter what order you have your account numbers listed. That is, you can group those who have participated and those who have not, or you can list them in order of account number, or any other way you'd like. However, you should indicate in some way what month and year these data are for. You can do this by adding another column beside the Account column, or by appending numbers to the account number.

Figure 2: Including Dates



	A	B
1	Account—date	
2	5495321-1/99	
3	5579382-1/99	
4	6293204-1/99	

Here dates are attached to account numbers.

	A	B
1	Account	Date
2	5495321	1/99
3	5579382	1/99
4	6293204	1/99

Here dates are in their own column.

Of course you can indicate dates in any way you want. You can use words (Jan. 99) or a different type of number (9901).

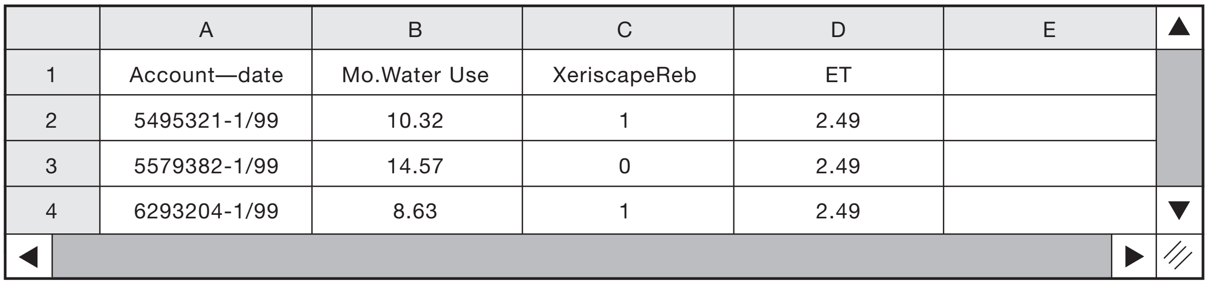
➨ Use Descriptive Variable Names

Be sure to use meaningful names for your variable names! You will have to use these names later (when you get to "Actually Doing the Analysis" and "Interpreting What You've Got"). At that point, you'll be sorry if you've named them Var1, Var2, or something equally uninformative.

All the other columns—for the other variables—have to be matched to each account number.

1. So, the next column would likely be consumption data. For each account, you enter the amount of water consumed in the next column, same row—all in the consumption column.
2. Now, suppose the third column is for your Xeriscape rebate program. Next to each account number would be a 1 for each account that participated in the Xeriscape rebate program that month and a 0 for each account that did not participate. If you organized all your participants together and then added non-participants, this column would look like a whole bunch of 1s followed by a whole bunch of 0s. If you organized it some other way—say, by address—then you would see a random scattering of 1s and 0s.
3. Suppose that the fourth column is ET (evapotranspiration) data and you have a small community, so only the data from one weather station are relevant to you. Then, this column would contain the same ET information for each account in the database. If your community was larger and you use data from two weather stations, this column would have a scattering of ET data from one station and a scattering of ET data from the other—always matched to (in the same row as) the appropriate account.

Figure 3: Adding Variables



	A	B	C	D	E
1	Account—date	Mo.Water Use	XeriscapeReb	ET	
2	5495321-1/99	10.32	1	2.49	
3	5579382-1/99	14.57	0	2.49	
4	6293204-1/99	8.63	1	2.49	

In this example, different accounts use different amounts of water (measured in units), 2 have received a Xeriscape rebate and 1 has not, and all are near the same weather station.

Matching Dichotomous Variables to Dates

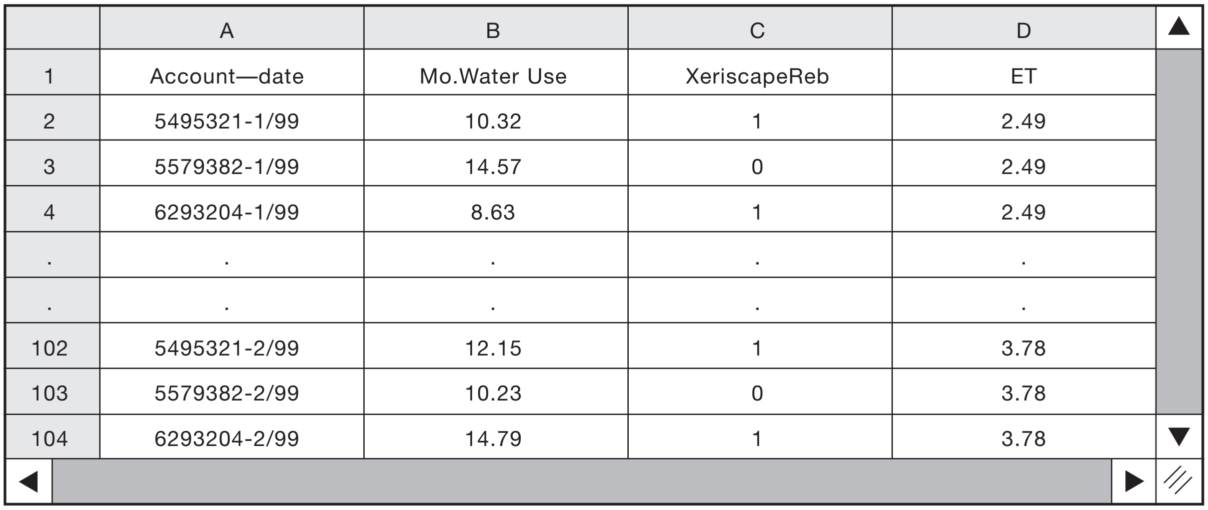
When you are using dichotomous variables to indicate whether an account has received a program or not (as discussed in item 2 and the figure immediately above), an issue that can arise is when a household that has received a program should start having a 1 as opposed to the 0 that indicates the program has not been received. To make this more concrete, suppose that a household participated in a Xeriscaping seminar in January. Would it get a 1 in the January row of the Xeriscape Seminar variable column?

Here's the principle: you don't want to measure a household as having received a program until the program actually could have made a difference. So, how do you decide? Well, if you have an exact date, then you could split the month, and count those who attended the seminar between January 1 and January 15 as having received the program in January, but count those who attended the seminar between January 15 and January 31 as having received it in February. After all, If the seminar was attended on January 25th, it's very unlikely the householder could possibly have time to save measurable water in January. If you don't have exact date information, then it is probably safest to measure the program as having been received in the *next* month—everyone who attended the seminar gets a 0 in the Xeriscape Seminar variable column for January and a 1 in the Xeriscape Seminar variable column for February. That way, when you measure a household as having received a program in a month, you know they have had it for the *whole* month.

Now, suppose that next month you go back in to add another month's data. This time, you are following the same people, but some things have changed while others have not.

1. Copy the column of account numbers and add them to the bottom of your original account column.
   - If you appended month and year numbers to your account numbers, don't forget to change these!
2. Now, in the Consumption column, enter the new consumption data for this month, again matching them to the account numbers.
   - Suppose that you have 100 accounts you are tracking.
   - Item 2 in your Account column (the first account number entered, but the title will be item 1) and item 102 in your Account column will be the same account number.
   - Item 2 in your Consumption column will be that account's consumption the first month, and item 102 in your Consumption column will be that account's consumption the second month.
3. For the Xeriscape Participation column, copy it and append it to the bottom of the original Xeriscape column—generally, whether these people participated or not has not changed in one month. If some of those who had not participated have now participated, change those 0s to 1s.
4. Go to the ET column, and enter the new ET data for this month. If you use only one weather station, now this column will have the same number in it 100 times, and then a new number repeated 100 times.

Figure 4: Adding Another Time Period



	A	B	C	D
1	Account—date	Mo.Water Use	XeriscapeReb	ET
2	5495321-1/99	10.32	1	2.49
3	5579382-1/99	14.57	0	2.49
4	6293204-1/99	8.63	1	2.49
.	.	.	.	.
.	.	.	.	.
102	5495321-2/99	12.15	1	3.78
103	5579382-2/99	10.23	0	3.78
104	6293204-2/99	14.79	1	3.78

In this example,

- Different accounts use different amounts of water and different households (accounts) use different amounts of water in different months;
- Which have received a Xeriscape rebate has not changed in one month;
- All are near the same weather station and ET has changed between the months of January and February.

Data Can Be Organized in Many Ways

The data don't have to be organized in just this way. Suppose you get your first year's data pulled for you by someone in billing records. In this case, they will probably be organized with the first account number repeated 12 times (one time for each month) followed by the next account number repeated 12 times, etc. This is fine, too. But here you need to note that then the ET data, for example, would have 12 different numbers followed by the same set of 12 different numbers, followed by the same set of 12 different numbers, etc. (for a single-weather-station community).

The principles here are (1) Be sure to store everything that is the same variable in the same column—all account numbers go in the account number column, no matter what order they go in, and all ET data go in the ET data column, etc.—and (2) Be sure that each entry for each variable is matched to the account number and time period in the same row.

What the Data Should Look Like

Now that you know how to enter data in the computer, you should also know what they should look like. In general, data for multivariate analysis should either be dichotomous (variables that take on only the values 0 or 1 where 0 means "no" and 1 means "yes") or continuous.

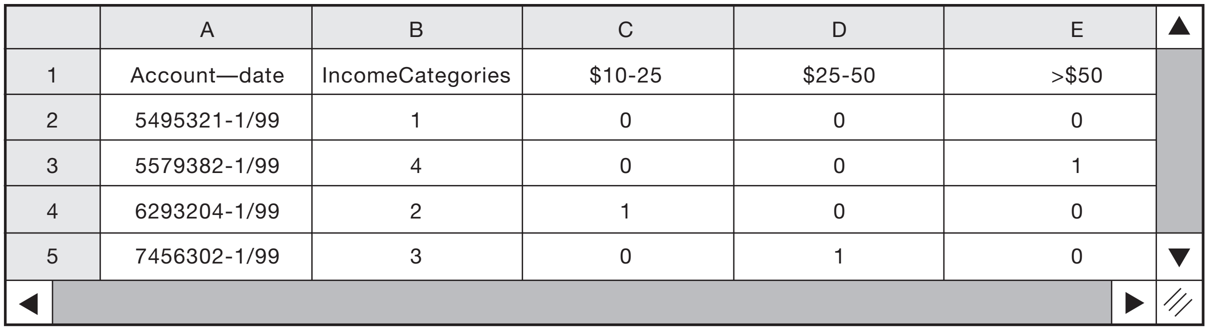
Dichotomous variables are sometimes called "dummy" variables—but it's not because they are dumb! Many things in real life are naturally dichotomous: are you male or female? Did you vote or not? Did you get a Xeriscape rebate, yes or no?

Continuous numbers include numbers that are measured in whole numbers (like total population) and also numbers that are measured in percents or decimals (like what percent of households have children, or how many units of water a household used this month).

Sometimes data are collected and stored differently. For example, consider household income level. Sometimes, categories of income are used (for example, "less than $10,000"; "between $10,000 and $25,000"; "More than $25,000 but less than $50,000"; "More than $50,000") and then data are stored in a computer as numbers that stand for the categories (for example, here 1 might stand for "less than $10,000," 2 for "between $10,000 and $25,000," 3 for "More than $25,000 but less than $50,000," and 4 for "More than $50,000").

Such data should not be used as is for multivariate analysis. The first lesson is, if you have the choice, don't collect data this way. Second, if you already have such data, then you need to recode them. You will need to create new variable columns. The number of columns should be one less than the number of categories you have. So, using the example given above, you will need 3 new variable columns. The first one will take on a 0 unless the account is category 2 ("between $10,000 and $25,000"), in which case it will take on a 1; the second one will take on a 0 unless the account is category 3, and then 1; and the third one will take on a 0 unless the account is category 4 and then 1.

Figure 5: Converting From Categorical to Dichotomous Variables



	A	B	C	D	E
1	Account—date	IncomeCategories	$10-25	$25-50	>$50
2	5495321-1/99	1	0	0	0
3	5579382-1/99	4	0	0	1
4	6293204-1/99	2	1	0	0
5	7456302-1/99	3	0	1	0

In this example, income data were collected in four categories and coded with numbers to indicate each category. The analyst created three new variables and converted the income data into those new variables.

Note that, taking the three new variables together, for each row at MOST one cell can have a 1 in it.

Similarly, if race/ethnic data are stored as (for example) 1 = White, 2 = Black, 3 = Hispanic, 4 = American Indian, 5 = Other, you will need to recode the data. Again, you will need as many new variables as one less than the number of categories, so you will need four new variable columns. Here the first variable column (labeled "Black") would be 0 unless the account-holder is black, and then 1; the second (labeled "Hispanic") would be 0 unless the account-holder is Hispanic, and then 1; the third would be 0 unless the account-holder is American Indian, and then 1; and the fourth would be 0 unless the account holder's racial/ethnic group is "Other," and then 1.

If you are using Census data, remember you'll use tract fractions for each group, not dichotomous variables.

⟹ Converting Variables is Easy

Converting one variable into several others as described here is relatively easy. It can be done either directly in Excel (using an "If" statement) or in SPSS or other statistical software. So, don't worry if you've originally coded your data in the categorical way. Just remember to create the new variables as outlined here *before* you run the analysis.

One reason people often collect sensitive data categorically or in ranges (as described above) is that some people may be reluctant to reveal, for example, their exact household income. However, you can ask people to point to the income level that is closest to theirs, with incomes listed in tens-of-thousands of dollars (be sure to include amounts clear up to amounts that cover at least 90% of your population). Alternatively, you can ask people to state their incomes to the nearest $10,000. Then, if you enter the dollars they point to, you can use these data directly since they are "continuous enough."

Be sure to have a reason that you can explain to people why they should reveal their household incomes, such as "we understand that income data can be sensitive, but we would really appreciate it because research indicates that how much water people use is partly determined by household income." Similarly, some people may not want to tell you their ages, but they may be willing to tell you the year they were born. Think creatively about how to measure data at the continuous level whenever possible.

Q&A: Incomplete Survey Data

Suppose I have survey data for some accounts and not for others?

Answer

Use the more precise data (the survey data) whenever you have them and fill in gaps with less precise data, such as Census data. For example, suppose some people refuse to answer the survey question about income. Use the income data for those who do answer and use the appropriate Census tract average for those who don't.

Structuring the Community-Level Data File

For the community-level data (for doing community-level analysis), you won't have account numbers. Instead, the "index column"—the first column, that tells you the unit of observation for each row—is simply the time period (date). Remember, it is best to collect data as frequently as possible, up to monthly, but quarterly or even yearly is still useful. The next column is total residential water consumption for the time period, matched to each date. The other variable columns will be similar to those for account-level data, but matched to the time-period. For example, if you are collecting yearly data, you will only be able to include a column for each water conservation measure that indicates (using a 1) each year the program was in effect and indicates (using a 0) each year the program was not in effect (to include a bit more information, you could use fraction of each year, in decimals, that each program was in effect, with 0 indicating none of the year and 1 indicating the whole year, and, for example, 0.5 indicating 6 months).

The hard part about the community-wide data file is getting the demographic data (data about your population and its composition) frequently enough. If all you are able to use is Census data, your data on, for example, population, will only change every 5 or 10 years. Usually, this is not often enough to be useful. So, you will need to find alternate data sources. Still, be sure to store what you DO have here. It can still come in handy.

Working With the Minimum

This section discusses the minimum data needs for the following types of small studies.

Suppose you don't have all the data recommended for your "Data Pantry," but you still want to perform some analysis. What are the minimum ingredients you need in order to produce something that has the potential to be useful for decision-making? The following tells you three minimum combinations for hurry-up analysis.

But, remember, using these minimum combinations decreases the likelihood that the results will be very useful or very accurate: a white-flour biscuit is worth eating, but it isn't as nutritious as a loaf of whole-grain bread and, if you get unlucky and the baking soda is bad, you won't want to eat it at all. The point here is that these studies have the potential to be useful, but you might get unlucky and find nothing at all, not because your conservation measures aren't useful, but because these studies using minimal data simply aren't good enough to detect the use.

And remember, for all of these studies, using more than the minimums described is better—as long as you're careful to meet the rules about the number of observations relative to the number of independent variables.

The Smallest Analysis

A Time-Series Analysis

FIGURE 6: A SMALL TIME-SERIES ANALYSIS

1. Computer with Excel, SPSS, and ~10 megabytes of free space, or the equivalent
2. Total residential water consumption for your community, preferably about 35 observations (for example, three years of data at the monthly level = 36 observations)
3. Start- and end-date information for all water conservation measures that were in effect during the study period (the study period is determined in item 2)
4. Population size for your community OR Total number of residential water accounts, for the years covered by your analysis
5. Climate data—ET and precipitation
6. Complete information on all residential water prices in effect during the years you're analyzing
7. Consumer Price Index (CPI) data for putting prices in constant dollars
8. Narrative discussing programs

If your Community has been changing over the study period

9. Median household income and Percent in poverty
10. Percent of population aged 17-24
11. Percent of households with children
12. Percent of population with HS diploma or above
13. Average house OR Account age
14. Average house value

Here your dependent variable would be community-wide residential water use, at least at the yearly level (more often is better). In addition, at the very least, you must have measures of the following:

- Observations for all water conservation programs that were in effect for some (but not all) of the study period,
  - preferably measured in dollars spent during each time at which you're measuring consumption or measured by percent of population that received the program at each time; or
  - (second-best) percent of each measurement-period in effect; or
  - (third-best) at least whether it was in effect during each time period or not;
- Total population or total residential accounts;
- Price, including sewer and other charges customers may view as part of the price (in real/constant/base-year dollars);
- ET and precipitation data; and
- Household income data, including median household income, and percent of population in poverty.

If your community has been changing importantly over the time-period to be analyzed, you should also include measures of the following, as possible:

- Average house/account age;
- Average house value (in real/constant/base-year dollars);
- Percent of population between the ages of 17 and 24;
- Percent of population with high school education or above;
- Percent of households with children; and
- Percent of houses with flood (off-meter) irrigation.

Variables Must Vary

All variables must vary over the time period to be analyzed, or they cannot be included. This is ALWAYS true.

How Much Data?

At the VERY MINIMUM, you MUST have more data on the dependent variable than you have variables in the model. So, if you have 5 independent variables, you must have AT LEAST 6 observations for the dependent variable (that is, 6 years of data on community-wide residential water use, or 6 months of data on community-wide residential water use).

Still, people are unlikely to believe your results if you don't have at least 5 more observations than independent variables.

And, that having been said, it is really best to have 30-33 more observations than independent variables. Therefore, with 5 independent variables, you should have 35-38 observations on the dependent variable. You're unlikely to have that much yearly data, but that would require only about three years of monthly data or about 7 or 8 years of quarterly data.

The Main Problem

The biggest problem with this type of analysis is that the level of aggregation is so great that you may not be able to detect any effect of your water conservation programs, especially if many of your programs were small and individually delivered (rather than large or community-wide). Also, as for all of the minimum studies described in this section, the estimates will not be as precise, and may be biased, compared to the full model.

Other Small Analyses

A Small Cross-Sectional Analysis

FIGURE 7: A SMALL CROSS-SECTIONAL ANALYSIS

1. Computer with Excel, SPSS, and 10-50 megabytes of free space, or the equivalent
2. Information that allows you to link program recipients to their water accounts
3. Water consumption data for some who participated and some who didn't, for the same time period
4. Climate data—data on ET and precipitation ONLY if your community is big enough that different accounts may use different weather station data
5. Data on household differences, including
   - Number of persons in the household
   - Whether there are children in the household
   - Number of people aged 17-24 in the household
   - Household income
   - Whether the household is Hispanic and/or whether English is spoken
   - Educational attainment
   - House age
   - House value
   - Number of bathrooms (or bedrooms)
   - Availability of flood (off-meter) irrigation

Suppose you've got consumption data for a group of households that participated in a water conservation program, but you don't have data for these households before they participated. In that case, you could do a cross-sectional study, where you compare those who have received the program with others who have not received it.

For this type of study, you need the following information

- Consumption data for other households that did not receive the conservation program, observed at the same time as for the group that did (that is, if your consumption data is for the participators in May of 2000, then you need to use consumption data for non-participators during May of 2000; if your data on participators is for the year of 1999, then your data for non-participators also needs to be for 1999).
- Climate data ONLY if your community is big enough that climate affects different households importantly differently at the same time period.
- Data on household differences, including the following:
  - Household income;
  - How many people are in each household;
  - How many people ages 17-24 are in each household;
  - Which households have children present;
  - Whether the household is Hispanic and/or whether English is spoken;
  - Educational attainment;
  - Value of each house;
  - Age of each house;
  - Which houses have flood irrigation; and
  - Number of bathrooms in each house.

REMEMBER, if you can't get these data house-by-house, you can use Census tract averages.

How Much Data?

The rule of thumb is that you need at least 33 more observations on the dependent variable than the number of independent variables. However, for this type of study, if participators are a small percent of your community's total population, you would *like* to have more non-participators than participators in the study, if possible. Try to have at least 30 participators and at least 50 non-participators (as usual, more, up to 1200, is better).

The Main Problem

The main problem with this study is that you don't know if there were differences in water consumption between those who chose to participate and those who didn't BEFORE participation. Maybe people chose to participate because their water bills were unusually high. In that case, your study might show participators using the same or even more water than non-participators, even though the water conservation program DID save water—it's just that these folks started out using more water to begin with. Alternatively, maybe people who are conservation-minded chose to participate, and they used less water before the program than did the others—so the effect you're seeing isn't really due to the program, but to conservation-mindedness. As for all of the minimum studies, the estimates will not be as precise, and may be biased, compared to the full model.

A Small Before-And-After Analysis

FIGURE 8: A SMALL BEFORE-AND-AFTER ANALYSIS

1. Computer with Excel, SPSS, and 10-50 megabytes of free space, or the equivalent
2. Information that allows you to link program recipients to their water accounts
3. Information on all conservation programs that people may have received during the study period (not just the one of primary interest)
4. Information on all prices that have been in effect, if more than one set
5. CPI data for deflating prices
6. Climate data—data on ET and precipitation
7. Data on household differences, including
    - Number of persons in the household
    - Whether there are children in the household
    - Number of people aged 17-24 in the household
    - Household income
    - Whether the household is Hispanic and/or whether English is spoken
    - Educational attainment
    - House age
    - House value
    - Number of bathrooms (or bedrooms)
    - Availability of flood (off-meter) irrigation

Here you would have an observation on the water consumption of some households/accounts before they participated in a water conservation program and after they participated. This is a type of pooled, or panel, study. If you've got only one or so observations before and one or so observations after, the observations should be relatively soon before and after—say, not more than one year before and not more than one year after.

For this type of study, you need the following, if they have changed during the time to be studied:

- Price measures, including sewer and other charges that customers will view as part of the price (in constant dollars);
- ET and precipitation (weather) measures; and
- Measures of other conservation programs that the households received between the first time the household is included and the last time it is included, including community-wide programs if they were in effect during part of the time observed and not during part.

In addition, for each household, you should try to have data on the following factors which probably won't change for the household before and after, but are different between households:

- Household income;
- How many people are in each household;
- How many people ages 17-24 are in each household;
- Which households have children present;
- Whether the household is Hispanic and/or whether English is spoken;
- Educational attainment;
- Value of each house;
- Age of each house;
- Which houses have flood irrigation; and
- Number of bathrooms in each house.

If you can't collect these data for each household, you should use Census tract averages to fill in.

How Much Data?

The rule of thumb is that you need at least 33 more observations on the dependent variable than the number of independent variables. So, if each household/account is observed once before it received the water conservation program and once after (2 observations of each household), if you include about 9 of the variables above, you need a minimum of 21 households. MORE (up to about 1200) IS ALWAYS BETTER!! More observations increase your confidence in the results.

The Main Problem

The main problem with this type of study is that you are only observing people who participated. If you chose participants randomly, this probably doesn't matter. But, if people are volunteers for a program, then you may not be able to generalize the results to your whole community—only to people who are willing to volunteer. As for all of the minimum studies, the estimates will not be as precise, and may be biased, compared to the full model.

Actually Doing the Analysis

This section discusses the steps in actually doing the analysis.

It also discusses:

With modern statistical software, actually doing the data analysis is usually very simple. It's getting ready to do the analysis and then interpreting what you've got that take time. If you've already gotten together a data pantry and at least enough data to perform one of the minimum analyses described in the earlier section, then you are almost ready to do analysis, but there are a few things you must check and do first.

1. Create a Data File of the Variables That You Actually Plan to Use for This Analysis

You can create this by cutting and/or pasting. For example, make a duplicate of your data file, and then delete all variables that you won't be using for this analysis.

Don't forget to create new dichotomous variables for any variables that have been entered and stored categorically (as discussed in the section "Organizing Your Data For Analysis").

DON'T USE MASTER DATA FILES FOR ANALYSIS

Your master data files should not be used directly for data analysis. You don't want to lose those data if something goes wrong!

2. Do Some File Checking to Make Sure You Can Use All the Data You Plan to Use

- **First, check each variable to make sure that it varies during the time period you will be analyzing this time. Delete every variable that does not vary for the study period.** Here, you are checking down each column of data.

  For example, let's suppose that you are analyzing one year of data and that your provider's water prices have not changed during that year. Then, you will not use the price data for this study—though you still want to save those data for another study. You will delete the price variable(s) from the data file to be analyzed.

- **Check that there are no gaps for any observation. Delete any observation that does not have data for every variable you will be including in this analysis.** Here, you are checking across each row of data.

  Suppose your analysis will be account-level, and will include (among others) data on consumption, on household income, and on number of persons in the household. For every account, you must delete any account that does not include any ONE of the variables in the model. If account number 4454387 has all data but you don't know how many people in the household—and you can't fill in the gap, from Census or other data—then 4454387 has to be eliminated from this analysis. If 4454388 has all information but household income, 4454388 has to be eliminated.

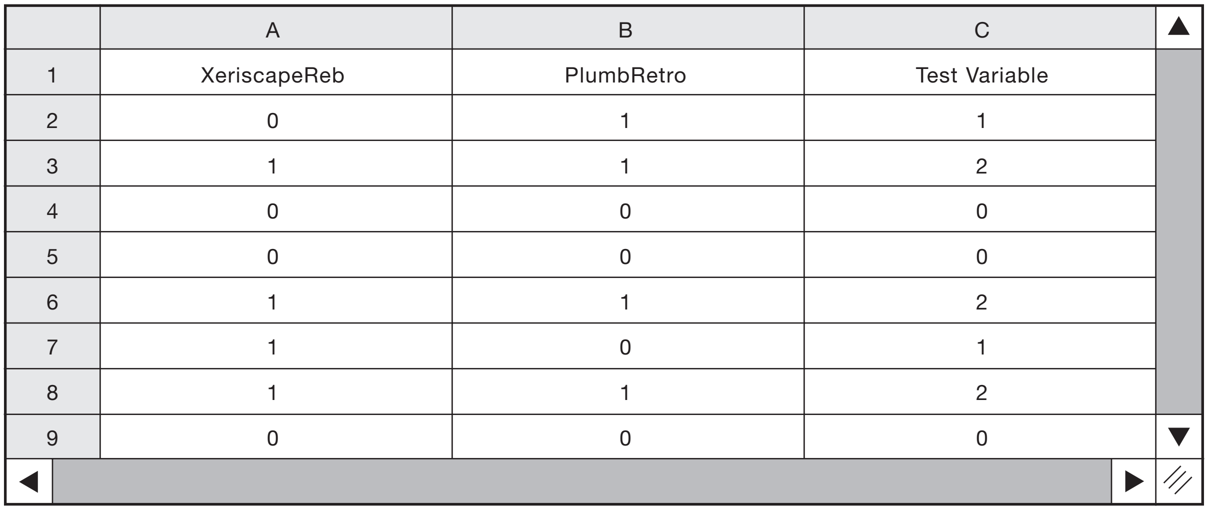
- **Check that numbers in the data are not too big or too small to make any sense.** Delete whole observations (whole rows) where there is one entry for one variable that is outside the possible range and where you cannot substitute the correct value.

3. Now That You Have Cleaned Your Data, You Need to Check Combinations of Dichotomous Variables

For reasons we won't bore you with, **it is very important that NO combination of the dichotomous variables can be summed so as to create an entire column that is nothing but 1s** (if this condition is violated, the computer will not be able to perform your statistical analysis correctly).

What does this mean?? Think of any one of your dichotomous variables—for example, whether an account has participated in the Xeriscaping Rebate Program or not. This variable is single column where every entry is either a 0 or a 1. Now, think about another dichotomous variable, such as whether account holders have received plumbing retrofits, also a column of 0s and 1s. Imagine adding these two variables together horizontally—that is, so that you are creating a new variable where each cell is the sum of the two values in the cells in the same row for the two previous variables. Call this horizontal sum "Test Variable."

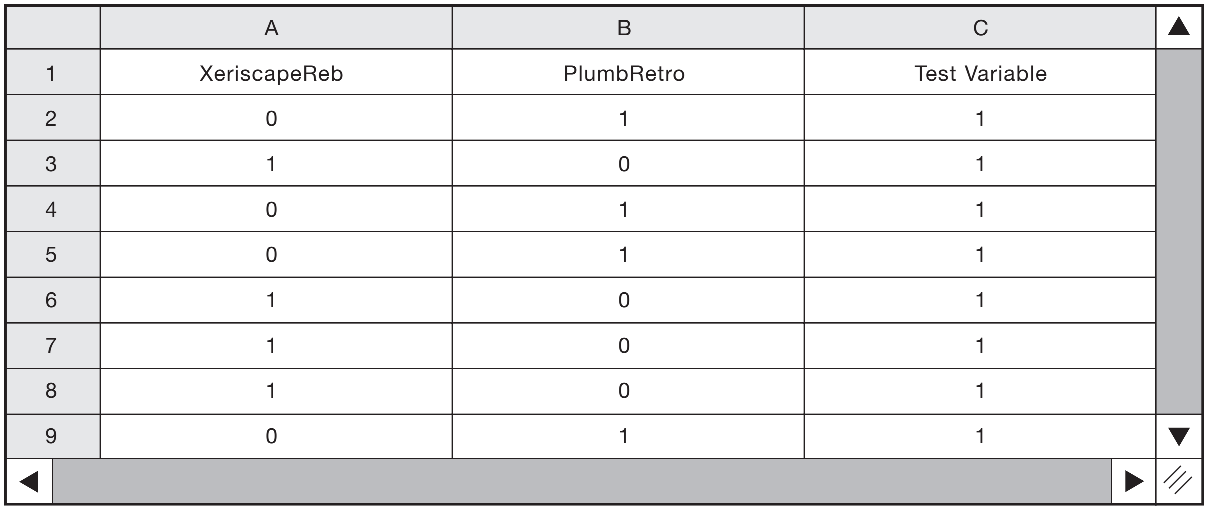
Figure 9: An OK Test Variable



	A	B	C
1	XeriscapeReb	PlumbRetro	Test Variable
2	0	1	1
3	1	1	2
4	0	0	0
5	0	0	0
6	1	1	2
7	1	0	1
8	1	1	2
9	0	0	0

Test Variable is not a whole column of 1s or the same multiple of 1, analysis can proceed.

Figure 10: A Problem Test Variable



	A	B	C
1	XeriscapeReb	PlumbRetro	Test Variable
2	0	1	1
3	1	0	1
4	0	1	1
5	0	1	1
6	1	0	1
7	1	0	1
8	1	0	1
9	0	1	1

Test Variable is a whole column of 1s, analysis cannot proceed.

If ANY COMBINATION of dichotomous variables can be summed to create an entire column (a new variable, such as "Test Variable") that is nothing but 1s, you must delete enough dichotomous variables (entire columns) until this is no longer true. Delete as few as you can and still meet this requirement.

4. If Your Study Period is Longer Than One Year, Use Your Price Index Data to Convert All Dollar Values to Constant/Base-year Dollars

To do this, first look at your price index data. Usually, price index data are given in numbers with two or three places before the decimal point, and two or three places after the decimal place (e.g., a price index number might be like 123.45 or 78.96). Sometimes, price index numbers are given as fractions, where values will either be slightly greater than 1 (like 1.23) or less than one (like 0.78).

- IF AND ONLY IF your numbers are given with two or three places before the decimal point—in the form XXX.XX—divide the column by 100. Then, divide each variable (column) that is measured in dollars by the new column (the original price index numbers divided by 100), to create variables that are measured in constant dollars. (These column operations can be done in Excel or in the statistical package).
- If the price index is given with only no or one number before the decimal place (in the X.XX form), then do NOT divide your price index data by 100. Directly divide each variable (column) that is measured in dollars by the price index column to create variables that are measured in constant dollars. Remember to name these new columns something you can remember.

5. Next, Convert All Continuous Variables into Logarithmic Functions of the Original Data but DO NOT TAKE THE LOG OF ZERO

This is simple to do. Simply take the logarithm to the base 10 or the natural logarithm of each column (variable), using either the function in Excel or the function in your statistical package. **Don't forget to take the log of the dependent variable, water consumption.** Only take the log of variables that will actually go in the analysis; don't take the log of index variables, like account number. **Remember, don't do this to the dichotomous variables.**

Why? The purpose of this conversion is to take into account the tapering-off effect known for the relationship between water consumption and other variables that cause it (for example, as wealth increases, water consumption increases, but not to the same extent when people are increasing household income from $100,000 per year to $110,000 as when they are increasing it from $20,000 per year to $30,000 per year).

DO NOT TAKE THE LOG OF ZERO!

The log of zero is undefined, so do not try to take the logarithm of zero. Alter continuous variables that can equal zero before taking the log. If you are taking the logarithm of continuous variables that may legitimately take on the value zero, then you need to convert the zeros to a number that is sufficiently small that you are willing to consider it as zero for purposes of the analysis. For example, suppose that in some households there are 0 people ages 17 through 24. Clearly, this is a legitimate number of people of that age (especially if you live in Sun City!). For this variable, which could vary between 0 and 6, or so, 0.4 is basically the same as zero, so substitute 0.4 for all zeros before you take the log (create a new variable that replaces all 0 values with 0.4). Similarly, suppose your water price was originally zero (at some point in the study period, you didn't charge for water). Depending on the price that you charged when you changed this policy, then perhaps $0.10 would be a reasonable 0 price.

6. Go Into Your Statistical Package, and Run a Correlation Matrix— If You Can Figure Out How

The correlations you care about appear "off the diagonal" of the correlation matrix. The "diagonal" is the stripe of 1.00s that you will see running in an angular line down the matrix. Examine the off-diagonal correlations to make sure that none of them are much over 0.8 (positive or negative). If they are, then there may be a problem when you run the analysis (see section, "Possible Problems When You Run the Program"). If any of them are 1 (or approximately 1, like 0.98) then there WILL be a problem when you run the analysis.

The following gives an example of a correlation matrix.

Figure 11: A Sample Correlation Matrix

	%Hispanic	XeriscapeReb	PlumbRetro	%WithChild
%Hispanic	1.00	0.26871	0.037839	0.46913
XeriscapeReb	0.26871	1.00	0.036511	0.40553
PlumbRetro	0.037839	0.036511	1.00	0.29956
%WithChild	0.46913	0.40553	0.29956	1.00

Note: Shading is used to stress the off-diagonal; actual correlation matrices won't includes shading.

Note the diagonal line of 1.00s. This is because the correlation of a variable with itself is always perfect, equal to 1. Also notice that the pieces of the correlation matrix that are off the diagonal—the side from the top down to the diagonal and the side from the bottom up to the diagonal—are mirror images of each other. This is because the correlation of two variables is the same no matter what order you do it. Because of this, some programs don't show one half of the correlation matrix. In this case, it would look as shown in Figure 12.

Figure 12: A Sample Correlation Matrix With Half Omitted

	%Hispanic	XeriscapeReb	PlumbRetro	%WithChild
%Hispanic	1.00			
XeriscapeReb	0.26871	1.00		
PlumbRetro	0.037839	0.036511	1.00	
%WithChild	0.46913	0.40553	0.29956	1.00

No matter how the information is presented, in this example the off-diagonal correlations are quite moderate (all 0.5 or less), indicating that there are no problems.

7. Go into Your Statistical Package, and Run a Regression Analysis Using Consumption as the Dependent Variable and All the Converted Variables and Dichotomous Variables as the Independent Variables

- Remember to use the logarithmed versions of the variables, and don't forget that you should have taken the logs of dollar variables that were converted to constant dollars, not the original ones.
- Don't forget to include the dichotomous variables that you have NOT taken the logs of.
- Don't get mixed up and include index variables—such as account numbers, or dates, or addresses— in the regression.

An Intermediate Technique

This intermediate technique involves the measurement of house age. As you may recall from the earlier discussion, it is expected that newer homes use less water (when everything else is the same) because their fixtures and appliances are more water-saving than those in older homes. On the other hand, it was stated earlier that this effect should "taper off." For the statistical analysis to be able to estimate this tapering effect, it is necessary to enter both the Home Age variable, as well as another variable that measures age. If you don't really care about home age, but are only including it as a control, you may not want to do this (though doing it will improve the overall quality of the model). If you care about home age—for example, if you have a program that will be targeted to older house—then you should do the following.

- Take your Home Age variable and use it to create an additional variable, Home-Age-Squared. Do this by squaring the Home Age variable (multiply Home Age by Home Age) using either Excel or your statistical package.
- You now have two variables, Home Age and Home-Age-Squared. Take the logarithm of both of these variables, and use them both when you run the analysis.

Advanced Techniques

If you have a statistical package that will allow you to do this, the following are some techniques that you can use to make your results more robust.

- If it is an option, instruct the program to use "White" or "Robust" or "Huber" Standard Errors when it runs the regression.
- If this is not an option and you are using cross-sectional or pooled data, consider controlling for heteroskedasticity (sometimes called "heteroscedasticity"), using "household income" as the control variable if you are asked for one.
- If you are performing a time-series analysis or a pooled analysis, and White, Robust, and Huber errors are not options, consider controlling for autocorrelation (perhaps referred to as autoregression or AR1). If you are asked, you want to control for first-order autocorrelation.
- Note that for the general type of pooled data, with people who have and have not received water conservation programs followed over time, you probably need to control for both heteroskedasticity and autocorrelation if White, Robust, or Huber standard errors are not available.

Possible Problems When You Run the Program

Sometimes, data on the independent variables are too highly correlated. When this happens, the statistical procedure performed by the statistical package cannot work properly. There are different ways that different packages alert you to this problem.

In SPSS, for example, the program will kick out one of the overly correlated variables before it runs the program. There will be a message to this effect in the output. In some other programs, you will only be able to tell this problem exists by examining the standard errors that are included in the program output. If any standard error (also referenced as "Std. Err." or "SE") is extremely large—say, 10 or more times the size of the coefficient estimate—then this is a warning sign, and you need to check the correlation of that variable with others. Even if you couldn't figure out how to produce a correlation matrix (in step 6, above), you should be able, fairly easily, to do correlations one-by-one with other variables. Start with those that are either similar to that variable (measured similarly, like other percent variables or other dichotomous variables) or that have little variation (like dichotomous variables, again, or perhaps your price variable).

Problem

The problem here is that you cannot understand the independent effect of the different variables that are highly correlated. Suppose two of your conservation programs are so highly correlated that they can't be used. You can go ahead and do the analysis using only one, but you can't tell the effect of that one: the estimates give you the effect of both mixed together, and you can't separate the effects.

Solution

There is only one solution to this problem: collect more data on at least one of the independent variables that are too highly correlated. The fundamental problem here is a lack of variability, underlining variability's importance (discussed previously in "Stocking Your Data Pantry").

INTERPRETING WHAT YOU'VE GOT

This section discusses how to find and use the results you have.

Once you run the regression, you will be faced with output including several tables filled with numbers that probably don't mean anything to you. Here, we discuss how to make those numbers make sense.

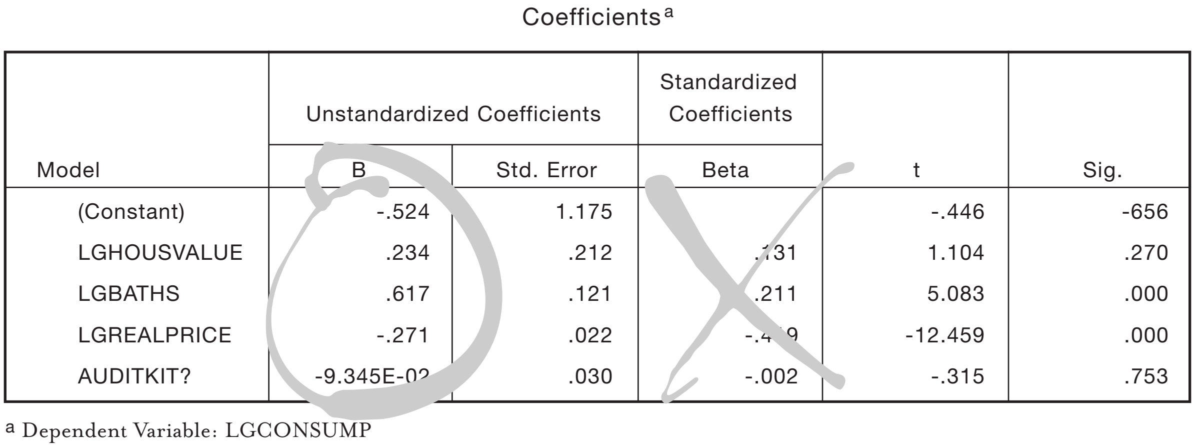
Finding the Results

The purpose of all this work has been to get the coefficient estimates that are produced by the regression package. So, the first thing you need to do is find them on the print-out. Unfortunately, this can be somewhat tricky.

In SPSS, the table you want will be several pages in. It is called "Coefficients" and the column you are interested in is called "B" under the heading "Unstandardized Coefficients." Do not use the numbers under the column "Beta" under "Standardized Coefficients." The reason this needs to be stated is that some programs may list the coefficient estimates under the heading "Beta" (in case you're wondering, the reason for this confusion is that what political scientists call "beta" and what economists call "beta" are not the same things). **If there is both something called "B" or "parameter estimate" or "coefficient estimate" and something called "Beta," don't use "Beta."** That having been said, if you are in doubt, you need to look in your statistical software documentation and find out what "Beta" is; if beta is defined as a *standardized* coefficient or parameter, don't use it (you want the *unstandardized* coefficient/parameter estimate).

Also, ignore the number beside the word "Constant" (or "Const," or something similar). Usually this will be either the first number reported in the table with coefficients/parameters or the last number.

FIGURE 13: SAMPLE OUTPUT (FROM SPSS)



Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.524	1.175		-.446	-656
LGHOUSVALUE	.234	.212	.131	1.104	.270
LGBATHS	.617	.121	.211	5.083	.000
LGREALPRICE	-.271	.022	-.419	-12.459	.000
AUDITKIT?	-9.345E-02	.030	-.002	-.315	.753

[a] Dependent Variable: LGCONSUMP

This figure shows a sample coefficient table from SPSS. Use the column labeled "B." Don't use the column of "Beta" figures, and ignore the first row "(Constant)."

Coefficient tables usually appear several pages into the output. Formatting may be somewhat different depending on what version of SPSS you are using, and formatting can be completely different for other statistical packages.

In general, the program will give you much more information and many more types of numbers than you can use or understand. Don't worry about it; just focus on those you can understand and that are useful to you. Using a highlighter to highlight those you want to work with can help.

Converting Estimates

Once you find the coefficient estimates (under the "B" heading for SPSS), you probably will need to convert some of the numbers. If the numbers are simply numbers followed by decimals, then no conversion is needed. However, if the numbers are followed by decimals and then also the letter "E" and some other stuff, then you need to convert the numbers. For example, an estimate could be -9.345. This number does not require conversion. However, if an estimate is -9.345E-02 (as for "AuditKit?" above), then it does require conversion. The "E-02" is a shorthand way of saying that the number is multiplied times 10 to the minus-two power, which is 0.01. This is called "scientific notation." In general, "E-0X" has "X" numbers after the decimal place, with the last number a 1. So, "E-05" would be 0.00001. The easiest way to convert the number is to move the decimal point to the left as many times as the number after the "E." So, if the number is -9.345E-02, you move the decimal over once, to get -0.9345, and then again to get -0.09345. If you don't find this way easy, you can figure out what the number "E-0X" equals (as E-02 equals 0.01) and then multiply that number times the number in front of the "E" (here, multiply -9.345 by 0.01).

Once you've converted all the numbers that are given in scientific notation, you are ready to begin interpreting the coefficient estimates. You interpret them one by one, and they tell you the relationship between EACH variable and the independent variable, if all the other variables stay the same (that is, this is the independent effect of each independent variable on the dependent variable).

There are two different ways to interpret them, depending on whether you took the log of the independent variable in question (for continuous variables) or did not (for dichotomous variables), and these two ways are discussed in the following sub-sections. But first, read the next note.

SOME COEFFICIENTS RELATE TO ALL WATER USERS, OTHERS TO A SUBSET

When you are interpreting the coefficients, keep in mind that some are relevant to many more users than are others. For example, a water price change will typically impact all residential water users. Therefore, savings can be spread over all residential users—once you've figured out the savings (below) then you can multiply the savings by all residential accounts. However, some programs are targeted to sub-groups. For example, if you have some type of "Seniors Helping Seniors" program, then you can only expect the savings for senior households that are willing to participate. So, you need to multiply household savings by a much smaller subset of your population to get total expected savings.

If the coefficients are around the same size, total savings will tend to be much less for targeted programs than for city-wide programs.

Interpreting Coefficients for Logarithmed Independent Variables

When both the dependent variable and the independent variable are measured in logarithms, then the coefficient estimate can be interpreted as an elasticity. Basically, an elasticity tells you the effect of a one-percent change in the independent variable on the dependent variable, also in percents.

- So, if the coefficient estimate for an independent variable measured in logs is 0.04, this says that a 1% increase in that independent variable results in a 0.04% increase in the dependent variable. If the coefficient is 1.1, then a 1% increase in the independent variable results in a 1.1% increase in the dependent variable. This also means that a 1% decrease in the independent variable results in a 1.1% decrease in the independent variable.
- If the coefficient value is positive, this tells you that, when one increases, the other increases; and when one decreases, the other decreases.
- If the coefficient value is negative, then it tells you the size of the *decrease* in the dependent variable from an increase in the independent variable. So, a coefficient of -1.1 indicates that a 1% *increase* in the independent variable results in a 1.1% *decrease* in the dependent variable. And, conversely, *decreasing* the independent variable by 1% results in an *increase* to the independent variable of 1.1%.

To interpret these percent changes further, you need to use a value for the dependent variable. If your dependent variable is community-wide residential water use for one month, then the first example (coefficient equal to 0.04) would predict an increase of 0.04% of total community residential water use. For coefficients with negative signs, to figure out how many gallons are saved, use either the most recent or the average community residential water use in a month, and multiply.

- Don't forget that a 1% increase in the independent variable may not be the most interesting increase. For example, suppose your water price is $1 per unit. It is unlikely you would ever raise the price by $0.01! Instead, you might raise it by $0.10. This is a 10% increase, so multiply the coefficient by 10 to get the effect on the dependent variable. For example, suppose that the coefficient for the price variable is -0.27 (as in the SPSS sample output, above). Then raising the price by $0.10, a 10% increase if the current price is $1, is predicted to decrease water consumption by 2.7%. If you might double water price, that is a 100% increase, so, in this example, you could predict a 27% decrease in total residential water consumption every month. However, you should note that you should be less confident of predictions that are big compared to changes that have occurred in your data: be less confident in the 27% change due to a dollar increase than in the 2.7% decrease due to a $.10 increase—unless your data include a 100% increase your provider has already tried.

If your dependent variable is water use by account, then use the mean (average) residential water use for your community as your base of prediction (it is easy to get either Excel or a statistical package to compute the average of any variable).

Go down every coefficient for a logarithmed independent variable that you care about (that are policy relevant), and write the meaning of the coefficient, both in percent terms and in terms of actual water saved. Make yourself a table, so that it will be clear to you what the coefficient estimates you have gained through all this hard work mean!

Coefficient Signs (Minuses and Pluses)

For water conservation programs and prices, you should expect—and hope!—that coefficient estimates will be negative. For these types of variables, a positive coefficient either means that the program actually *increases* water use, or else that it has no effect. You must exercise judgment to decide which. If the number implies a small positive effect on the dependent variable—or a medium-sized positive effect simply goes against your judgment and intuition—interpret the true effect as zero. But keep an open mind: there is reason to believe that some conservation programs can back-fire and cause water-users to stop being careful and end up using more water.

For other variables, coefficient estimates can be positive or negative. You may be able to predict ahead of time. For example, as the number of members of a household increase, the amount of water used will increase, if everything else is the same. For some, you may not have any way to predict. For example, would renters use more water than owners? It probably depends on who pays for renters' water, and you usually don't know that.

Interpreting Coefficients for Dichotomous Independent Variables

Remember that you did not take the logarithm of dichotomous variables. Also remember that these variables indicate the presence or absence of the variable characteristic. For example, a dichotomous variable might indicate that a household did or did not receive a Xeriscaping rebate (though measuring the dollar amount of the rebate would be better, if you have the information). Or, a dichotomous variable might indicate whether a household is or is not Hispanic. Therefore, dichotomous variables are interpreted in terms of what happens when the characteristic is present relative to the situation when it is absent.

Because the dichotomous variables are not in logarithmic form, you do not interpret the effect of a 1-*percent* change on the dependent variable, but of a 1-*unit* change—that is, going from not having the characteristic to having it. Still, the effect of the 1-unit change is interpreted in terms of a *percent* change on the *dependent* variable. So, let's suppose that your community has a program of Seniors Helping Seniors with household audits and hardware retrofits, and you've measured this with a dichotomous variable, with a 1 for all accounts that received this program. Next, let's suppose that your estimated coefficient is -0.06. This indicates that, if a household had not had Seniors Helping Seniors and now they do, they would use 6% less water than before. If you've measured monthly water use, this is 6% less water every month—times 12 months to get yearly savings. If you've measured yearly water use, this is 6% less water each year. **The coefficient is always interpreted in terms of the way that the dependent variable is measured.**

As with logarithmed variables, positive coefficients estimated for dichotomous variables indicate that, when the attribute is present, the dependent variable increases relative to when the attribute is absent. Negative ones indicate that the dependent variable decreases when the attribute is present.

For Dichotomous Variables Where There are Multiple Categories (Conversions of Categorical Variables)

In earlier sections, we talk about converting categorical variables into dichotomous variables. An example that was used was for race/ethnicity. How do you interpret these coefficients? Recall that, for race/ethnicity, you were advised to use four variables, one for Hispanic, one for Black, one for American Indian, and one for Other. This means that there is not an explicit variable for White/Non-Hispanic. In this type of case, each dichotomous variable is interpreted *relative to* the base category, which is the category that does not have an explicit variable. In this example, the base category is White/Non-Hispanic. Suppose that you had this type of data for race, and your coefficient estimates worked out to be Hispanic = +0.1, Black = -0.0001, American Indian = -0.1, Other = +0.0001. This indicates that American Indians use less water than White/Non-Hispanics, Hispanics more than other whites, and Blacks and Others about the same. More specifically, you would interpret this as indicating that, if a household was Hispanic instead of White/Non-Hispanic, but everything else about that household was the same, then you would expect it to use 10% more water every time period. (You might care about this if you wanted to target your conservation materials or create programs for specific ethnic groups.)

Just as another example, to make sure the point is clear, if you have household income data in various category variables, then, again, you interpret the coefficients relative to the base category—the one that doesn't have its own variable. If you don't have an explicit variable for the poorest category (as in the example earlier), then you interpret the coefficients for each other category relative to the poorest households.

If You Used the Intermediate Technique (Home Age and Home-Age-Squared)

If you included two variables for home age, both Home Age and Home-Age-Squared, then interpreting the coefficients for these two variables is somewhat different than for other coefficients in the analysis. First, if the hypothesis that newer houses are more water-efficient than older houses is supported, you should expect the coefficient for the logarithm of Home Age to be positive (as the house gets older, more water is used). Second, if it is also true that the effect of age tapers off as older fixtures and devices are replaced, then you should expect the coefficient of the logarithm of Home-Age-Squared to be negative. Then, if you want to interpret the meaning of these coefficients, you need to calculate the following formula:

- (Coefficient of Home Age) + (Coefficient of Home-Age-Squared)(2)(Average Logarithm of Home Age)

Use the Average Logarithm of Home Age from your own data (take the average or mean of your Logarithmed Home Age variable in either Excel or your statistical package). Where one value in parenthesis is touching another value in parenthesis—as for (2)(Average Logarithm of Home Age)—multiply the values together. The result will tell you the overall effect of a one-percent increase in home age on water consumption, in percents, taking into account both the age of fixtures and replacement of them.

If You Used Advanced Techniques

The following information is relevant ONLY if you used the techniques discussed in the "Advanced Techniques" section of "Actually Doing the Analysis."

If You Were Able to Use "White" or "Robust" or "Huber" Standard Errors

Then you may also use the information contained in the "*t*" and the "R-squared" statistics (also referred to as "R Square," "R2," "R^2"), as described below.

White, Robust, and Huber Standard Errors

If White, Robust, or Huber standard errors are options, they should correct for both heteroskedasticity and autocorrelation. However, this may not always be true, and they may control for only one of these conditions. In that case, the following information is not correct for pooled data—or for the type of condition they do not control for. For example, if they control for heteroskedasticity and you are using pooled or cross-sectional data, then do not use the *t* statistic or R-squared information.

The *t* Statistic

The *t* statistic is a measure of statistical significance. Basically, it relates to how likely it is that the true value of each coefficient you've estimated is really 0 (usually, you would like this to be *unlikely*.) Most statistical programs automatically generate a *t* statistic for each coefficient estimated. In general, bigger *t* statistics (in magnitude, ignoring the sign) indicate that it is less likely that the true value of the coefficient is 0. Anything over 1.96 is "statistically significant at the 95% confidence level or higher." Numbers larger than -1.96 in magnitude (such as -2.00) are also "statistically significant at the 95% confidence level or higher"—the *t* statistic simply takes the same sign as the coefficient estimate, so just ignore the sign.

All this having been said, there are two points to remember.

1. *t* values smaller than 1.96 can also indicate sufficient levels of statistical significance, depending on how many observations you have in your data set and also how confident you need to be to make policy decisions. Most people don't need to be 95% confident they will win before they'll make a bet. Would you bet on an 80% probability? How about 75%? More specifically, are you willing to spend public money on a water conservation program if you are 60% confident that it is saving water?
   - Many programs that compute *t*s also provide the significance level (perhaps labeled "Sig."). Interpreting this column can be a little tricky. Often, small numbers in the significance column are better than large ones. To see if this is true for your program, find a *t* statistic that is close to or bigger than 1.96. Then, see if the significance is close to 0.95 or close to 0.05. If the latter, then smaller "Sig."s indicate higher confidence; if the former, then larger "Sig."s indicate higher confidence. For SPSS, small "Sig." numbers indicate high levels of statistical confidence. If you look at the sample output (Figure 13), you'll see that the "Sig." for the *t* that is much larger than 1.96 (5.083) is .000, indicating a very high level of statistical confidence (at the .999 level) that the true value of the coefficient is not zero.
2. Even if you have a small *t* statistic, your best bet is NOT that the true value of the coefficient is actually 0. Your best bet is still the value of the coefficient. If your coefficient is 0.1 and your *t* statistic is 1.05, for example, you are better off betting that the true value of the coefficient is 0.1 than you are betting that the true value is 0.

The most valuable use of the *t* statistic is to create confidence intervals. Some packages (such as SPSS) can create these for you if you choose that option. In SPSS, a table called "Coefficients" "95% Confidence Interval for B" can be produced. If you find a water conservation program that is estimated to save water and where the confidence interval does not cross zero (both the "Lower Bound" and the "Upper Bound" are negative), then this is a VERY good bet—put money here.

The R-squared Statistic

The R-squared is a measure of how much of the variation in the dependent variable is explained by the independent variables included in the analysis. It can take on decimal values between 0 and 1, with higher numbers indicating that more of the variation in the dependent variable is explained. For example, an R-squared of 0.89 indicates that 89% of the variation in the dependent variable is explained by the independent variables.

It is a common error to believe that the R-squared is a measure of the quality of your model. This is not true in any simple way. First, you can always increase your R-squared by increasing the number of independent variables, holding constant sample size. In fact, a "good" way to get a high R-squared is to perform the small time-series analysis described in "Working With the Minimum." This especially works if you don't use very many time periods. Try having 10 time periods and 5 independent variables, and you can be sure of a high R-squared. BUT this model is much worse than a pooled model with, say, 30 independent variables and 1200 accounts. But the pooled model's R-squared will generally be lower. So, why bother talking about the R-squared?

- The R-squared can be useful in some ways:
  - If the R-squared is very high (above 0.9), this may be a sign of extreme correlation (as discussed before).
  - If the R-squared is very low—below 0.1, for example—you should be cautious in using the results.
  - Many people want to know the R-squared, so you should be prepared both to provide it and to understand it reasonably well.
  - If you are comparing VERY SIMILAR studies of similar phenomena using similar data with similar-sized datasets and a similar number of variables, bigger R-squareds may indicate more explanatory power. If independent variables are different in number, you may prefer to look at the "Adjusted R-squared," which is interpreted the same way as the R-squared but takes into account the number of independent variables. The Adjusted R-squared will usually be lower than the R-squared (it cannot be higher).

If You Are Using Time-Series Data, and You've Corrected for Autocorrelation
In this case, you may use the *t* and R-squared statistics as described above.

If You Are Using Cross-Sectional Data, and You've Corrected for Heteroskedasticity
In this case, you may use the *t* and R-squared statistics as described above.

If You Are Using Pooled Data and Corrected for Both Heteroskedasticity and Autocorrelation
You can use the *t* and R-squared statistics the same as if you have used Robust, Huber, or White standard errors (as described in that section, above).

If You Are Using Pooled Data and Have Corrected for Heteroskedasticity But Not For Autocorrelation
You can use the Durbin-Watson statistic to test for autocorrelation and, if it is not found, then you can use the *t* and R-squared statistics as described above (if it is found, you cannot use them). However, interpreting the Durbin-Watson statistic is not straightforward, and you will probably need to get some help to interpret it.

If You Are Using Pooled Data and Have Corrected for Autocorrelation But Not For Heteroskedasticity
There is no simple test for heteroskedasticity, so you cannot use the *t* and R-squared statistics.

Q&A: What if You Can't correct for Both

In pooled or panel data, is it worth correcting for either heteroskedasticity or autocorrelation if I can't correct for both?

Answer

Yes. Even though correcting for, say, heteroskedasticity when you can't correct for autocorrelation means you can't use the *t* and R-squared statistics, correcting is still valuable because it makes your coefficients more accurate. More accuracy means that you can be more confident that, for example, when you say raising the price by $0.10 will decrease water consumption by 2.7%, the effect really is close to that amount. If your statistical package lets you do it reasonably easily, it is worth it.

Conclusion

If you have gone through this book and done your best to follow its instructions, you are now ready to make decisions about which water conservation programs to keep spending resources on and which you should probably stop. You are also ready to make a report and/or presentation explaining what you've found. We're sure you know how to do that, so in closing we'll wish you good luck and leave you with two reminders:

1. Remember that water conservation program evaluation should be an ongoing process. Keep stocking your data pantry, and keep using the information you can gain from analysis to improve your water conservation efforts. See " ✏ DON'T Be Overwhelmed; DO Be Creative" if you need reinvigoration.
2. An advanced statistician will be able to tell you that there are a lot of weaknesses in the study you have done, whatever study you've done, but don't let that stymie you. It is better to eat a sandwich than to go hungry because you don't have time to fix a gourmet meal or the money to buy one. The analyses you can do following these guidelines are not the best analyses in the world, but they WILL give you reasonable information for decision-making until the perfect analysis comes along.

Glossary

Autocorrelation/Autoregression/AR1/Serial Correlation

This is a condition, common in time-series data, where the dependent variable at one time is related to the dependent variable at the next time period (or, at the previous time period). For example, consider measurement over time of the fraction of households in your community with children. The amount of households with children during the next period will be comprised mostly of the same households as last period, plus some new households with children, and minus some households where the children have gone to college or the household has left. This type of situation would cause (positive, first-order) autocorrelation.

Base-Year Dollars

See *Real/Constant/Base-Year Dollars*

Categorical Variables

Variables that use numbers to indicate categories (for example, using 1 to indicate White, 2 to indicate Black, 3 to indicate Hispanic, etc.), though the numbers themselves are not numerically meaningful. Categorical variables must be converted to Dichotomous Variables before being used in the analyses described here.

Coefficient

The coefficients are what are estimated by the regression procedure. They express the effect of a change in each independent variable, alone, on the dependent variable. We say that it is the effect of each independent variable *alone* to indicate that it is the effect if all the other independent variables remain fixed—it's the independent effect. Another way to think of the coefficient is to realize that it is the *slope* of the relationship between the dependent variable and each independent variable.

Constant Dollars

See *Real/Constant/Base-Year Dollars*

Continuous Variables

Variables measured as numbers with actual numeric meaning, whether it is units of water or percent of households with children or thousands of dollars. Continuous variables may be used in the analyses described here.

Correlated

Moving together. When variables are correlated, they have a tendency to move together. This can happen in two ways. In one type of correlation, when one variable increases, the other also tends to increase. An example of this type of correlation would be total water consumption and total accounts. In the other type of correlation, when one variable increases, the other tends to decrease. An example of this type of correlation would be rainfall and water consumption. Correlation between the dependent variable and the independent variables is desirable. Too much correlation between independent variables can cause problems.

Cross-Sectional Data/Analysis

Data or Analysis that includes many different individual entities, whether they are accounts or communities, observed at one point in time. Generally, for residential water conservation, cross-sectional analysis is the second-most-useful type, with pooled the most useful.

Data

Useful information! Not just numbers to be used in analysis, but also general information and context that helps you make sense of the numbers.

Dependent Variable

What you're trying to explain. Here, water consumption.

Dichotomous Variables

Variables that take on only the values 0 or 1. Examples might include gender, or whether an account has participated in a particular water conservation program or not. Dichotomous variables may be used in the analyses described here.

ET

See *Evapotranspiration*

Evapotranspiration (ET)

A measure of the amount of heat and evaporation occurring in the local environment. See AZMET, the Arizona Meteorological Network, at *Ag.Arizona.Edu/AZMET*.

Heteroskedasticity (also, Heteroscedasticity)

This is a condition that is common in cross-sectional data when the cross-sections are different in ways that mean that the dependent variable could vary a lot more for some of them than for others. For example, suppose that you were estimating a model of water consumption for water providers in the Phoenix AMA. It seems clear that Phoenix's water consumption could vary a lot more than could Sun City's, simply because Sun City is much smaller than Phoenix. Similarly, think of the food budget for a poor household and a rich one. The poor household's food budget probably wouldn't vary much from month to month, but the rich one's might vary dramatically, depending on whether the household is just feeding itself or throwing a fancy party.

Independent Variables

Explanatory factors; factors that cause the dependent variable. Some are important to you—such as your conservation programs—some aren't, but need to be included so that you better understand the effect of the ones that are important.

Observation

The unit of analysis, for example, accounts. Generally, these are the rows in your data file.

Panel Data/Analysis

See *Pooled Data/Analysis*

Pooled Data/Analysis

Data or Analysis that observes more than one individual entity over more than one time period. For example, if you observe 100 accounts for 12 months, you are observing several entities (accounts) over several time periods (months), resulting in pooled data. Generally, pooled data are the most useful.

Proxy

A proxy variable is a variable that is used in place of the actual variable of interest when you can't observe what you really care about. To be a good proxy variable, the proxy must be highly correlated (see entry in this Glossary, above) with the variable of true interest. For example, the number of babies in a house is highly correlated with the number of diapers used, so the number of babies could be used as a proxy for the amount of diapers used if you can't directly observe the number of diapers used.

Real/Constant/Base-Year Dollars

Dollars (e.g., prices) that have been controlled for inflation, using some standard price index, such as the CPI (Consumer Price Index).

Standard Errors (SEs)

The Standard Error is a statistic that is used to compute both *t* statistics and R-squared statistics. Fundamentally, the standard error is a measure of the precision of the coefficient estimate. Therefore, you want each SE to be small compared to its own coefficient.

Study Period

The duration of time covered by a study. For a cross-sectional analysis, this could be only a single month or year. For a cross-time or pooled analysis, this could vary dramatically, from two months to 10 or more years. In short, the study period is from the first date you are observing data in this study to the last date you are observing data in this study.

Time-Series Data/Analysis

Data or Analysis of a single entity (usually a community) over several periods of time. For estimating the effectiveness of residential water-conservation programs, this is the least useful of the three types of data and analysis discussed in this cookbook.

Units (of Water)

A "unit" of water is equal to 748 gallons.

Variables

Factors that change over time or across accounts. Generally, these are the columns in your data file. Variables will measure factors that you care about—such as which accounts received a water program—as well as factors that you don't care about but which you need to include to control for other things that cause water consumption. Examples may include household income, or whether a household includes children or not.