7-2020

Quantitative Research Methods for Political Science, Public Policy and Public Administration for Undergraduates: 1st Edition With Applications in Excel

Wesley Wehde
East Tennessee State University, wehdew@etsu.edu

Tracey Bark

Hank Jenkins-Smith

Joseph Ripberger

Gary Copeland

See next page for additional authors

Recommended Citation

Wehde, Wesley; Bark, Tracey; Jenkins-Smith, Hank; Ripberger, Joseph; Copeland, Gary; Nowlin, Matthew; Hughes, Tyler; Fister, Aaron; and Davis, Josie. 2020. *Quantitative Research Methods for Political Science, Public Policy and Public Administration for Undergraduates: 1st Edition With Applications in Excel*. Johnson City, TN: East Tennessee State University.
https://dc.etsu.edu/etsu-oer/4

Authors
Wesley Wehde, Tracey Bark, Hank Jenkins-Smith, Joseph Ripberger, Gary Copeland, Matthew Nowlin, Tyler Hughes, Aaron Fister, and Josie Davis

FIRST EDITION

QUANTITATIVE RESEARCH METHODS FOR POLITICAL SCIENCE, PUBLIC POLICY AND PUBLIC ADMINISTRATION **FOR UNDERGRADUATES**

With Applications in **Excel**

WESLEY WEHDE

TRACEY BARK

HANK JENKINS-SMITH

JOSEPH RIPBERGER

GARY COPELAND

MATTHEW NEWLIN

TYLER HUGHES

AARON FISTER

JOSIE DAVIS

Quantitative Research Methods for Political Science, Public Policy and Public Administration for Undergraduates: 1st Edition With Applications in Excel

Wesley Wehde

Tracey Bark

Hank Jenkins-Smith

Joseph Ripberger

Gary Copeland

Matthew Nowlin

Tyler Hughes

Aaron Fister

Josie Davis

Preface and Acknowledgments

The idea for the graduate level version of this book grew over decades of teaching introductory and intermediate quantitative methods classes for graduate students in Political Science and Public Policy at the University of Oklahoma, Texas A&M, and the University of New Mexico. Despite adopting (and then discarding) a wide range of textbooks, we were frustrated with inconsistent terminology, misaligned emphases, mismatched examples and data, and (especially) poor connections between the presentation of theory and the practice of data analysis. The cost of textbooks and the associated statistics packages for students seemed to us to be, frankly, outrageous. So, we decided to write our own book that students can download as a free PDF, and to couple it with R, an open-source (free) statistical program, and data from the Meso-Scale Integrated Socio-geographic Network (M-SISNet), a quarterly survey of approximately 1,500 households in Oklahoma that is conducted with support of the National Science Foundation (Grant No. IIA-1301789). Readers can learn about and download the data here.

The idea of the undergraduate level of this book floated about amongst these various individuals until Fall 2019 when now Professor Wehde used the graduate level text in his undergraduate research methods course. He realized that, at times, the language of the text was at a higher level than necessary to introduce undergraduates in Political Science to research methods and statistics. This new version of the text omits large portions of the original text that focused on calculus and linear algebra, expands and reorganizes the content on the software system by shifting to Excelnd includes guided study questions at the end of each chapter. He also proposed shifting the software over to Microsoft Excel which, while not free, is present in almost every workplace in 2020 when the book was revised. This version reflects that shift to Excel.

By intent, this book represents an open-ended group project that changes over time as new ideas and new instructors become involved in teaching graduate and the undergraduate methods in the University of Oklahoma Political Science Department and beyond. The first edition of the book grew from lecture notes and slides that Hank Jenkins-Smith used in his methods classes. The second edition was amended to encompass material from Gary Copeland's introductory graduate methods classes. The fourth (and a half) edition (this one!) was updated by Wesley Wehde, who currently manages and uses the book in his introductory quantitative methods courses for undergraduates in the East Tennessee State Political Science Department. The development of this version of the text was supported by an OER Award from the Sherrod Library at ETSU, as well. At this stage, Tracey Bark, now a professor at Auburn University Montgomery, was brought on as a co-author due to her expertise in using Excel in Research Methods.

In addition to instructors, the graduate assistants who co-instruct the methods courses are an essential part of the authorship team. The tradition started with Dr. Matthew Nowlin, who assisted in drafting the first edition in . Dr. Tyler Hughes and Aaron Fister were instrumental in implementing the changes for the second edition. Dr. Wesley Wehde was responsible for much of the third and 4.5 edition and Josie Davis did most of the work on the fourth edition.

Copyright

CHAPTER ONE: Theories and Social Science

The focus of this book is on using quantitative empirical research to test hypotheses and build theory in political science and public policy. Quantitative means the book focuses on research that relies on data that can be quantified, or represented by numbers, as opposed to qualitative, or represented primarily by words. Empirical means this text focuses on research that involves measuring phenomenon in the real world using the scientific method as opposed to anecdotes or other types of evidence. Testing hypotheses and building theory means this text focuses on research that uses logic, and statistical techniques, to arrive at reasonable conclusions about the world or potential states of the world.

The book is designed to be used by undergraduate students in introductory courses to research methods, statistics, and quantitative analysis in the social sciences. It is important to note that quantitative analysis is not the only – or even the most important – kind of analysis undertaken in political science and public policy research. Qualitative analysis, including ethnographic studies, systematic cases analyses, focus groups, archival studies, and qualitative elite interviews (to name only a few approaches) are of critical importance for understanding social and political phenomena. With that understanding in mind, this book and the associated courses focus on the development and application of systematic analysis, hypothesis testing and theory building using quantitative data and modeling. Specifically, we focus on developing research design, univariate analysis, and a basic understanding of linear regression modeling and analysis. Throughout we provide applications and examples using the Excel statistical platform.

The Scientific Method

Empirical research, as outlined in this book, is based on the scientific method. Science is a particular way that someepistemologists believe we can understand the world around us. Science, as a method, relies on both logic, as captured by theory, and empirical observation of the world to determine whether the theory we have developed conforms to what we actually observe. We seek to explain the world with our theories, and we test our theories by deducing and testing hypotheses. When a **working hypothesis** is supported, we have more confidence in our theory. When the **null hypothesis** is supported, it undermines our proposed theory.

Science seeks a particular kind of knowledge and has certain biases. When we are engaging in scientific research we are interested in reaching generalizations. Rather than wanting to explain why President Trump's approval dropped, we are interested in explaining why presidential approval drops across various presidents, or, better yet, how economic conditions affect presidential approval. These generalizations should be logical (which is nothing more than saying they should be grounded in a strong theory) and they should be empirically verified (which, we will see means that we have tested hypotheses deduced from our theory). We also look for generalizations that are causal in nature. Scientists actively seek explanations grounded in causation rather than correlation. Scientific knowledge should be replicable – meaning that other scholars should be able to reach the same conclusions that you do. There should be inter-subjective agreement on scientific findings – meaning that people, with different personal experiences and biases, should still reach the same conclusion.

Scientists also tend to prefer simple explanations to complex ones. They have a bias that says the world is pretty simple and that our theories should reflect that belief. Of course, people are complex, so in the social sciences it can be dangerous to look only for the simplest explanation as most concepts we consider have multiple causes.

Theory and Empirical Research

This book is concerned with the connection between theoretical claims and empirical data. It is about using statistical modeling; in particular, the tool of regression analysis, which is used to develop and refine theories. We define **theory** broadly as a set of interrelated propositions that seek to explain and, in some cases, predict an observed phenomenon.

Theory: A set of interrelated propositions that seek to explain and predict an observed phenomenon.

Theories contain three important characteristics that we discuss in detail below.

Characteristics of Good Theories

- Coherent and internally consistent
- Causal in nature
- Generate testable hypotheses

Coherent and Internally Consistent

The set of interrelated propositions that constitute a well structured theory are based on **concepts**. In well-developed theories, the expected relationships among these concepts are both coherent and internally consistent. **Coherence** means the identification of concepts and the specified relationships among them are logical, ordered, and integrated. An **internally consistent** theory will explain relationships with respect to a set of common underlying causes and conditions, providing for consistency in expected relationships (and avoidance of contradictions). For systematic quantitative research, the relevant theoretical concepts are defined such that they can be measured and quantified. Some concepts are relatively easy to quantify, such as the number of votes cast for the winning Presidential candidate in a specified year or the frequency of arrests for gang-related crimes in a particular region and time period. Others are more difficult, such as the concepts of democratization, political ideology or presidential approval. Concepts that are more difficult to measure must be carefully **operationalized**, which is a process of relating a concept to an observation that can be measured using a defined procedure. For example, political ideology is often operationalized through public opinion surveys that ask respondents to place themselves on a Likert-type scale of ideological categories.

Concepts and Variables

A concept is a commonality across observed individual events or cases. It is a regularity that we find in complex world. Concepts are our building blocks to understanding the world and to developing theory that explains the world. Once we have identified concepts we seek to explain them by developing theories based on them. Once we have explained a concept we need to define it. We do so in two steps. First, we give it a dictionary-like definition, called a nominal definition. Then, we develop an operational definition that identifies how we can measure and quantify it.

Once a concept has been operationalised and possibly quantified, it is employed in modeling as a **variable**. In statistical modeling, variables are thought of as either dependent or independent variables. A **dependent variable**, Y, is the outcome variable; this is the concept we are trying to explain and/or predict. The **independent variable(s)**, X, is the variable(s) that is used to predict or explain the dependent variable. The expected relationships between (and among) the variables are specified by the theory.

Measurement

When measuring concepts, the indicators that are used in building and testing theories should be both **valid** and **reliable**. Validity refers to how well the measurement captures the concept. Face validity, for example, refers to the plausibility and general acceptance of the measure, while the domain validity of the measure concerns the degree to which it captures all relevant aspects of the concept. Reliability, by contrast, refers to how consistent the measure is with repeated applications. A measure is reliable if, when applied to the repeated observations in similar settings, the outcomes are consistent.

Assessing the Quality of a Measure

Measurement is, in quantitative research, the process of assigning numbers to the phenomenon or concept that you are interested in. Measurement is straight-forward when we can directly observe the phenomenon. One agrees on a metric, such as inches or pounds, and then figures out how many of those units are present for the case in question. Measurement becomes more challenging when you cannot directly observe the concept of interest. In political science and public policy, some of the things we want to measure are directly observable: how many dollars were spent on a project or how many votes the incumbent receives, but many of our concepts are not observable: is issue X on the public's agenda, how successful is a program, or how much do citizens trust the president. When the concept is not directly observable the operational definition is especially important. The operational definition explains exactly what the researcher will do to assign a number for each subject/case.

In reality, there is always some possibility that the number assigned does not reflect the true value for that case, i.e., there may be some error involved. Error can come about for any number of reasons, including mistakes in coding, the need for subjective judgments, or a measuring instrument that lacks precision. These kinds of error will generally produce inconsistent results; that is, they reduce reliability. We can assess the reliability of an indicator using one of two general approaches. One approach is a test-retest method where the same subjects are measured at two different points in time. If the measure is reliable the correlation between the two observations should be high. We can also assess reliability by using multiple indicators of the same concept and determining if there is a strong inter-correlation among them using statistical formulas such as Cronbach's alpha or Kuder-Richardson Formula 20 (KR-20).

We can also have error when our measure is not valid. Valid indicators measure the concept we think they are measuring. The indicator should both converge with the concept and discriminate between the concept and similar yet different concepts. Unfortunately there is no failsafe way to determine whether an indicator is valid. There are, however, a few things you can do to gain confidence in the validity of the indicator. First, you can simply look at it from a logical perspective and ask if it seems like it is valid. Does it have face validity? Second, you can see if it correlates well with other indicators that are considered valid, and in ways that are consistent with theory. This is called construct validity. Third, you can determine if it works in the way expected, which is referred to as predictive validity. Finally, we have more confidence if other researchers using the same concept agree that the indicator is considered valid. This consensual validity at least ensures that different researchers are talking about the same thing.

Measurement of Different Kinds of Concepts

Measurement can be applied to different kinds of concepts, which causes measures of different concepts to vary. There are three primary **levels of measurement**; ordinal, interval, and nominal. **Ordinal level** measures indicate relative differences, such as more or less, but do not provide equal distances between intervals on the measurement scale. Therefore, ordinal measures cannot tell us *how much* more or less one observation is than another. Imagine a survey question asking respondents to identify their annual income. Respondents are given a choice of five different income levels: $0-20,000, $20,000-50,000, $50,000-$100,000, and $100,000+. This measure gives us an idea of the rank order of respondents' income, but it is impossible for us to identify consistent differences between these responses. With an **interval level** measure, the variable is ordered and the differences between values are consistent. Sticking with the example of income, survey respondents are now asked to provide their annual income to the nearest ten thousand dollar mark (e.g., $10,000, $20,000, $30,000, ect.). This measurement technique produces an interval level variable because we have both a rank ordering and equal spacing between values. Ratio scales are interval measures with the special characteristic that the value of zero (0) indicates the absence of some property. A value of zero (0) income in our example may indicate a person does not have a job. Another example of a ratio scale is the Kelvin temperature scale, because zero (0) degrees Kelvin indicates the complete absence of heat. Finally, a **nominal level** measure identifies categorical differences among observations. Numerical values assigned to nominal variables have no inherent meaning, but only differentiate one ``type" (e.g., gender, race, religion) from another.

Theories and Causality

Theories should be causal in nature, meaning that an independent variable is thought to have a causal influence on the dependent variable. In other words, a change in the independent variable *causes* a change in the dependent variable. Causality can be thought of as the ``motor" that drives the model and provides the basis for explanation and (possibly) prediction.

The Basis of Causality in Theories

1. Time Ordering: The cause precedes the effect, $X \rightarrow Y$
2. Co-Variation: Changes in X are associated with changes in Y
3. Non-Spuriousness: There is not a variable Z that causes both X and Y

To establish causality we want to demonstrate that a change in the independent variable is a necessary and sufficient condition for a change in the dependent variable (though more complex, interdependent relationships can also be quantitatively modeled). We can think of the independent variable as a treatment, τ, and we speculate that τ causes a change in our dependent variable, Y. The ``gold standard'' for casual inference is an experiment where a) the level of τ is controlled by the researcher and b) subjects are randomly assigned to a treatment or control group. The group that receives the treatment has outcome Y_1 and the control group has outcome Y_0; the treatment effect can be defined as $\tau = Y_1 - Y_0$. Causality is inferred because the treatment was only given to one group, and since these groups were randomly assigned other influences should wash out. Thus the difference $\tau = Y_1 - Y_0$ can be attributed to the treatment.

Given the nature of social science and public policy theorizing, we often can't control the treatment of interest. For example, our case study in this text concerns the effect of political ideology on views about the environment. For this type of relationship, we cannot randomly assign ideology in an experimental sense. Instead, we employ statistical controls to account for the possible influences of confounding factors, such as age and gender. Using multiple regression we control for other factors that might influence the dependent variable.[1]

Generation of Testable Hypothesis

Theory building is accomplished through the testing of hypotheses derived from theory. In simple form, a theory implies (sets of) relationships among concepts. These concepts are then operationalized. Finally, models are developed to examine how the measures are related. Properly specified hypotheses can be tested with empirical data, which are derived from the application of valid and reliable measures to relevant observations. The testing and re-testing of hypotheses develops levels of confidence that we can have for the core propositions that constitute the theory. In short, empirically grounded theories must be able to posit clear hypotheses that are testable. In this text, we discuss hypotheses and test them using relevant models and data.

As noted above, this text uses the concepts of political ideology and views about the environment as a case study in order to generate and test hypotheses about the relationships between these variables. For example, based on popular media accounts, it is plausible to expect that political conservatives are less likely to be concerned about the environment than political moderates or liberals. Therefore, we can pose the **working hypothesis** that measures of political ideology will be systematically related to measures of concern for the environment – with conservatives showing less concern for the environment. In classical hypothesis testing, the working hypothesis is tested against a **null hypothesis**. A null hypothesis is an implicit hypothesis that posits the independent variable has no effect (i.e., null effect) on the dependent variable. In our example, the null hypothesis states ideology has no effect on environmental concern.

Theory and Functions

Closely related to hypothesis testing in empirical research is the concept of functional relationships – or functions. Hypotheses posit systematic relationships between variables, and those relationships are expressed as functions. For example, we can hypothesize that

[1] This matter will be discussed in more detail in the multiple regression section.

an individual's productivity is related coffee consumption (productivity is a *function* of coffee consumption).[2]

Functions are ubiquitous. When we perceive relational order or patterns in the world around us, we are observing functions. Individual decisions about when to cross the street, whether to take a nap, or engage in a barroom brawl can all be ascribed to patterns (the ``walk" light was lit; someone stayed up too late last night; a Longhorn insulted the Sooner football team). Patterns are how we make sense of the world, and patterns are expressed as functions. That does not mean the functions we perceive are always correct, or that they allow us to predict perfectly. However, without functions we don't know what to expect; chaos prevails.

In mathematical terms, a function relates an outcome variable, y, to one or more inputs, x. This can be expressed more generally as: $y = f(x_1, x_2, x_3, \ldots x_n)$, which means y is `a function of the x's, or, y varies as a function of the x's.

Functions form the basis of the statistical models that will be developed throughout the text. In particular, this text will focus on linear regression, which is based on linear functions such as $y = f(x) = 5 + x$, where 5 is a constant and x is a variable. This type of function is the basis of the linear models we will develop, therefore these models are said to have a **linear functional form**.

However, non-linear functional forms are also common. For example, $y = f(x) = 3 - x^2$ is a quadratic function, which is a type of polynomial function since it contains a square term (an exponent). This function is non-linear because the changes in y are not consistent across the full range of x.

Examples of Functions in Social Science Theories

As noted, functions are the basis of statistical models that are used to test hypotheses. Below are a few examples of functions that are related to social science theories.

- Welfare and work incentives
  - Employment $= f$(welfare programs, education level, work experience,…)
- Nuclear weapons proliferation
  - Decision to develop nuclear weapons $= f$(perceived threat, incentives, sanctions,…)
- ``Priming'' and political campaign contributions
  - Contribution(\$) $= f$(Prime (suggested \$), income,…)
- Successful program implementation

[2] The more coffee, the greater the productivity – up to a point! Beyond some level of consumption, coffee may induce the jitters and ADD-type behavior, thereby undercutting productivity. Therefore the posited function that links coffee consumption to productivity is non-linear, initially positive but then flat or negative as consumption increases.

- Implementation $= f$(clarity of law, level of public support, problem complexity,...)

Try your hand at this with theories that are familiar to you. First, identify the dependent and independent variables of interest; then develop your own conjectures about the form of the functional relationship(s) among them.

Theory in Social Science

Theories play several crucial roles in the development of scientific knowledge. Some of these include providing patterns for data interpretation, linking the results of related studies together, providing frameworks for the study of concepts, and allowing the interpretation of more general meanings from any single set of findings. Hoover and Todd (2004) provide a very useful discussion of the role of theories in ``scientific thinking" – find it and read it!

The Role of Theory in Social Science

Adapted from *The Elements of Social Scientific Thinking* by Kenneth Hoover and Todd Donovan (2004, 37)

- Theory provides patterns for the interpretation of data
- Theory links one study with another
- Theory supplies frameworks within which concepts acquire significance
- Theory allows us to interpret the larger meaning of our findings

Perhaps, in the broadest sense, theories tie the enterprise of the social (or any) science together, as we build, revise, criticize and destroy theories in that collective domain referred to as ``the literature."

Outline of the Book

The goal of this text is to develop an understanding of how to build theories by testing hypotheses using empirical data and statistical models. There are three necessary ingredients of strong empirical research. The first is a carefully constructed theory that generates empirically testable hypotheses. Once tested, these hypothesis should have implications for the development of theory. The second ingredient is quality data. The data should be valid, reliable, and relevant. The final ingredient is using the appropriate model design and execution. Specifically, the appropriate statistical models must be used to test the hypotheses. Appropriate models are those that are properly specified, estimated, and use data that conforms to the statistical assumptions. This course focuses on model design and execution.

As noted, this text uses political ideology and views on the environment as a case study to examine theory building in the social sciences.[3] The text is organized by the idealized steps

[3] As you may have already realized, social scientists often take these steps out of order … we may ``back into" an insight, or skip a step and return to it later. There is no reliable cookbook for what we do. Rather, think of the idealized steps of the scientific process as an

of the research process. As a first step, this first chapter discussed theories and hypothesis testing, which should always be (but often are not!) the first consideration. The second chapter focuses on research design and issues of internal and external validity. Chapter 3 examines data collection methods. Chapter 4 covers specific ways to understand how the variables in the data are distributed. This is vital to know before doing any type of statistical modeling. The sixth chapter covers inference and how to reach conclusions regarding a population when you are studying a sample. The seventh chapter explores how to understand basic relationships that can hold between two variables including cross tabulations, covariance, correlation, and difference of means tests. These relationships are the foundation of more sophisticated statistical approaches and therefore understanding these relationships is often a precursor to the later steps of statistical analysis. The eighth through tenth chapters focus on bivariate ordinary least squares (OLS) regression, or OLS regression with a dependent variable and one independent variable. This allows us to understand the mechanics of regression before moving on the third section (chapters eleven to thirteen) that cover multiple OLS regression.

As a final note, this text makes extensive use of `Excel`. The steps to reproduce all of the examples is included in the text in such a way that readers should be able to replicate the results themselves. The data used for the examples is available as well. You can find it here.

Study Questions

1) What are the three necessary components of well-constructed, empirical research?
2) What will be the case study used throughout this book?
3) Identify dependent and independent variables of interest to you; then develop your own conjectures about the form of the functional relationship(s) among them.
4) What factor denotes a ratio level of measurement as a subset of interval measurements?
5) Define null hypothesis.

CHAPTER TWO: Research Design

Research design refers to the plan to collect information to address your research question. It covers the set of procedures that are used to collect your data and explain how your data will be analyzed. Your research plan identifies what type of design you are using. Your plan should make clear what your research question is, what theory or theories will be considered, key concepts, your hypotheses, your independent and dependent variables, their operational definitions, your unit of analysis, and what statistical analysis you will use. It should also address the strengths and weaknesses of your particular design. The

important heuristic that helps us think through our line of reasoning and analysis – often after the fact – to help us be sure that we learned what we *think* we learned from our analysis.

major design categories for scientific research are *experimental designs* and *observational designs*. The latter is sometimes referred to as a correlational research design.

Overview of the Research Process

Often scholars rely on data collected by other researchers and end up, de facto, with the research design developed by the original scholars. But if you are collecting your own data this stage becomes the key to the success of your project and the decisions you make at this stage will determine both what you will be able to conclude and what you will not be able to conclude. It is at this stage that all the elements of science come together. We can think of research as starting with a problem or a **research question** and moving to an attempt to provide an answer to that problem by developing a theory. If we want to know how good (empirically accurate) that theory is we will want to put it to one or more tests. Framing a research question and developing a theory could all be done from the comforts of your backyard hammock. Or, they could be done by a journalist (or, for that matter, by the village idiot) rather than a scientist. To move beyond that stage requires more. To test the theory, we deduce one or more hypotheses from the theory, i.e., statements that should be true if the theory accurately depicts the world. We test those hypotheses by systematically observing the world—the empirical end of the scientific method. It requires you to get out of that hammock and go observe the world. The observations you make allow you to accept or reject your hypotheses, providing insights into the accuracy and value of your theory. Those observations are conducted according to a plan or a research design.

Internal and External Validity

Developing a research design should be more than just a matter of convenience (although there is an important element of that which we will discuss at the end of this chapter). Not all designs are created equally and there are trade-offs we make when opting for one type of design over another. The two major components of an assessment of a research design are its internal validity and its external validity. **Internal validity** basically means we can make a causal statement within the context of our study. We have internal validity if, for our study, we can say our independent variable caused our dependent variable. To make that statement we need to satisfy the conditions of causality we identified previously. The major challenge is the issue of **spuriousness**. We have to ask if our design allows us to say our independent variable makes our dependent variable vary systematically as it changes and that those changes in the dependent variable are not due to some third or extraneous factor, often called an ommitted variable. It is worth noting that even with internal validity, you might have serious problems when it comes to your theory. Suppose your hypothesis is that being well-fed makes one more productive. Further suppose that you operationalize "being well-fed" as consuming twenty Hostess Twinkies in an hour. If the Twinkie eaters are more productive than those who did not get the Twinkies you might be able to show causality, but if your theory is based on the idea that "well-fed" means a balanced and healthy diet then you still have a problematic research design. It has internal validity because what you manipulated (Twinkie eating) affected your dependent variable, but that conclusion does not really bring any enlightenment to your theory.

The second basis for evaluating your research design is to assess its **external validity**. External validity means that we can generalize the results of our study. It asks whether our findings are applicable in other settings. Here we consider what population we are interested in generalizing to. We might be interested in adult Americans, but if we have studied a sample of first-year college students then we might not be able to generalize to our target population. External validity means that we believe we can generalize to our (and perhaps other) population(s). Along with other factors discussed below, replication is a key to demonstrating external validity.

Major Classes of Designs

There are many ways to classify systematic, scientific research designs, but the most common approach is to classify them as experimental or observational. **Experimental designs** are most easily thought of as a standard laboratory experiment. In an experimental design the researcher controls (holds constant) as many variables as possible and then assigns subjects to groups, usually at random. If randomization works (and it will if the sample size is large enough, but technically that means infinite in size), then the two groups are identical. The researcher then manipulates the experimental treatment (independent variable) so that one group is exposed to it and the other is not. The dependent variable is then observed. If the dependent variable is different for the two groups, we can have quite a bit of confidence that the independent variable caused the dependent variable. That is, we have good internal validity. In other words, the conditions that need to be satisfied to demonstrate causality can be met with an experimental design. Correlation can be determined, time order is evident, and spuriousness is not a problem—there simply is no alternative explanation.

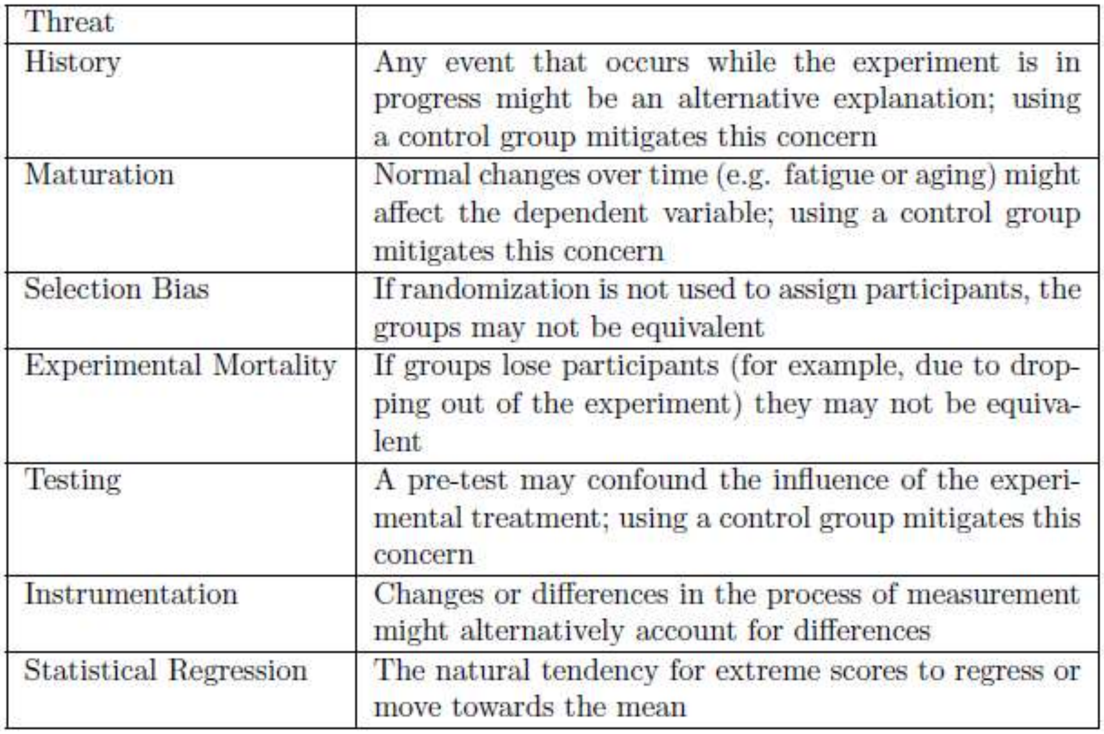
Unfortunately, in the social sciences the artificiality of the experimental setting often creates suspect external validity. We may want to know the effects of a news story on views towards climate change so we conduct an experiment where participants are brought into a lab setting and some (randomly selected) see the story and others watch a video clip with a cute kitten. If the experiment is conducted appropriately, we can determine the consequences of being exposed to the story. But, can we extrapolate from that study and have confidence that the same consequences would be found in a natural setting, e.g., in one's living room with kids running around and a cold beverage in your hand? Maybe not. A good researcher will do things that minimize the artificiality of the setting, but external validity will often remain suspect.

Observational designs tend to have the opposite strengths and weaknesses. In an observational design, the researcher cannot control who is exposed to the experimental treatment; therefore, there is no random assignment and there is no control. Does smoking cause heart disease? A researcher might approach that research question by collecting detailed medical and lifestyle histories of a group of subjects. If there is a correlation between those who smoke and heart disease, can we conclude a causal relationship? Generally the answer to that question is ``no", because any other difference between the two groups is an alternative explanation (meaning that the relationship might be spurious). For better or worse, though, there are fewer threats to external validity (see below for more detail) because of the natural research setting.

A specific type of observational design, the **natural experiment**, requires mention because they are increasingly used to great value. In a natural experiment, subjects are exposed to different environmental conditions that are outside the control of the researcher, but the process governing exposure to the different conditions arguably resembles random assignment. Weather, for example, is an environmental condition that arguably mimics random assignment. For example, imagine a natural experiment where one part of New York City gets a lot of snow on election day, whereas another part gets almost no snow. Researchers do not control the weather, but might argue that patterns of snowfall are basically random, or, at the very least, exogenous to voting behavior. If you buy this argument, then you might use this as natural experiment to estimate the impact of weather conditions on voter turnout. Because the experiment takes place in natural setting, external validity is less of a problem. But, since we do not have control over all events, we may still have internal validity questions.

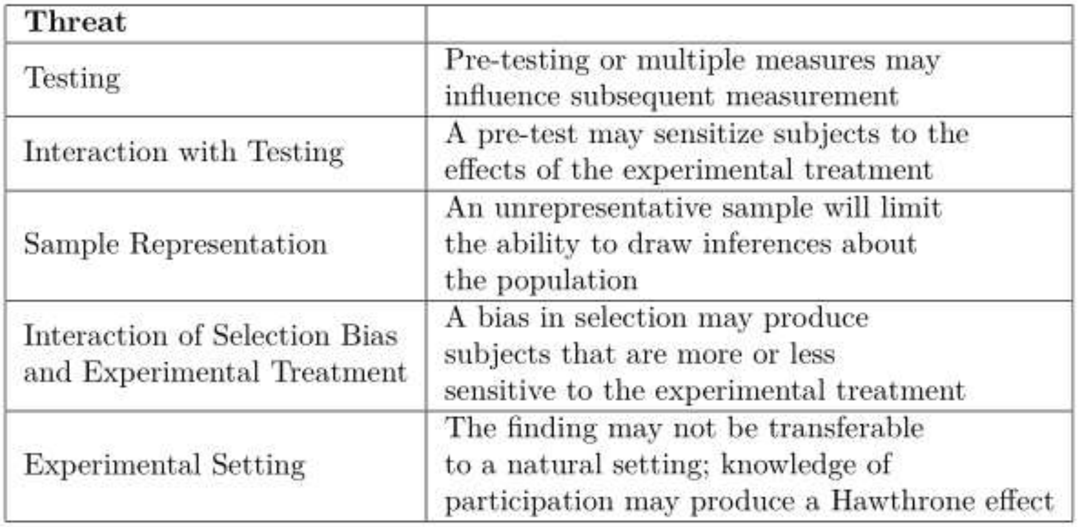
Threats to Validity

To understand the pros and cons of various designs and to be able to better judge specific designs, we identify specific **threats to internal and external validity**. Before we do so, it is important to note that a (perhaps ``the") primary challenge to establishing internal validity in the social sciences is the fact that most of the phenomena we care about have multiple causes and are often a result of some complex set of interactions. For examples, X may be only a partial cause of Y, or X may cause Y, but only when Z is present. Multiple causation and interactive affects make it very difficult to demonstrate causality, both internally and externally. Turning now to more specific threats, Table @ref(fig:tbl1) identifies common threats to internal validity and Table @ref(fig:tbl2) identifies common threats to external validity.



Threat	
History	Any event that occurs while the experiment is in progress might be an alternative explanation; using a control group mitigates this concern
Maturation	Normal changes over time (e.g. fatigue or aging) might affect the dependent variable; using a control group mitigates this concern
Selection Bias	If randomization is not used to assign participants, the groups may not be equivalent
Experimental Mortality	If groups lose participants (for example, due to dropping out of the experiment) they may not be equivalent
Testing	A pre-test may confound the influence of the experimental treatment; using a control group mitigates this concern
Instrumentation	Changes or differences in the process of measurement might alternatively account for differences
Statistical Regression	The natural tendency for extreme scores to regress or move towards the mean

Common Threats to Internal Validity



Threat	
Testing	Pre-testing or multiple measures may influence subsequent measurement
Interaction with Testing	A pre-test may sensitize subjects to the effects of the experimental treatment
Sample Representation	An unrepresentative sample will limit the ability to draw inferences about the population
Interaction of Selection Bias and Experimental Treatment	A bias in selection may produce subjects that are more or less sensitive to the experimental treatment
Experimental Setting	The finding may not be transferable to a natural setting; knowledge of participation may produce a Hawthrone effect

Common Threats to External Validity

Some Common Designs

In this section we look at some common research designs, the notation used to symbolize them, and then consider the internal and external validity of the designs. We start with the most basic experimental design, the post-test only design Figure (fig:post). In this design subjects are randomly assigned to one of two groups with one group receiving the experimental treatment.[4] There are advantages to this design in that it is relatively inexpensive and eliminates the threats associated with pre-testing. If randomization worked the (unobserved) pre-test measures would be the same so any differences in the observations would be due to the experimental treatment. The problem is that randomization could fail us, especially if the sample size is small.

$$\begin{array}{lll} R & X & O_1 \\ R & & O_2 \end{array}$$

Post-test Only (with a Control Group) Experimental Design

Many experimental groups are small and many researchers are not comfortable relying on randomization without empirical verification that the groups are the same, so another common design is the Pre-test, Post-test Design (Figure (fig:prepost)). By conducting a pre-test, we can be sure that the groups are identical when the experiment begins. The disadvantages are that adding groups drives the cost up (and/or decreases the size of the groups) and that the various threats due to testing start to be a concern. Consider the example used above concerning a news story and views on climate change. If subjects were given a pre-test on their views on climate change and then exposed to the news story, they might become more attentive to the story. If a change occurs, we can say it was due to the story (internal validity), but we have to wonder whether we can generalize to people who had not been sensitized in advance.

$$\begin{array}{llll} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \end{array}$$

Pre-test, Post-Test (with a Control Group) Experimental Design

A final experimental design deals with all the drawbacks of the previous two by combining them into what is called the Solomon Four Group Design (Figure @ref(fig:solomon)). Intuitively it is clear that the concerns of the previous two designs are dealt with in this design, but the actual analysis is complicated. Moreover, this design is expensive so while it may represent an ideal, most researchers find it necessary to compromise.

[4] The symbol R means there is random assignment to the group. X symbolizes exposure to the experimental treatment. O is an observation or measurement.

$$\begin{array}{llll} R & & X & O_1 \\ R & & & O_2 \\ R & O_3 & X & O_4 \\ R & O_5 & & O_6 \end{array}$$

Solomon Four Group Experimental Design

Even the Solomon Four Group design does not solve all of our validity problems. It still likely suffers from the artificiality of the experimental setting. Researchers generally try a variety of tactics to minimize the artificiality of the setting through a variety of efforts such as watching the aforementioned news clip in a living room-like setting rather than on a computer monitor in a cubicle or doing jury research in the courthouse rather than the basement of a university building.

Observational designs lack random assignment, so all of the above designs can be considered observational designs when assignment to groups is not random. You might, for example, want to consider the affects of a new teaching style on student test scores. One classroom might get the intervention (the new teaching style) and another not be exposed to it (the old teaching style). Since students are not randomly assigned to classrooms it is not experimental and the threats that result from selection bias become a concern (along with all the same concerns we have in the experimental setting). What we gain, of course, is the elimination or minimization of the concern about the experimental setting.

A final design that is commonly used is the repeated measures or longitudinal research design where repeated observations are made over time and at some point there is an intervention (experimental treatment) and then subsequent observations are made (Figure @ref(fig:repmeas)). Selection bias and testing threats are obvious concerns with this design. But there are also concerns about history, maturation, and mortality. Anything that occurs between O_n and O_{n+1} becomes an alternative explanation for any changes we find. This design may also have a control group, which would give clues regarding the threat of history. Because of the extended time involved in this type of design, the researcher has to concerned about experimental mortality and maturation.

$$O_1 \; O_2 \; O_3 \; O_n \;\; X \;\; O_{n+1} \; O_{n+2} \; O_{n+3}$$

Repeated Measures Experimental Design

This brief discussion illustrates major research designs and the challenges to maximizing internal and external validity. With these experimental designs we worry about external validity, but since we have said we seek the ability to make causal statements, it seems that a preference might be given to research via experimental designs. Certainly we see more and more experimental designs in political science with important contributions. But, before we dismiss observational designs, we should note that in later chapters, we will provide an approach to providing statistical controls which, in part, substitutes for the control we get with experimental designs.

Plan Meets Reality

Research design is the process of linking together all the elements of your research project. None of the elements can be taken in isolation, but must all come together to maximize your ability to speak to your theory (and research question) while maximizing internal and external validity within the constraints of your time and budget. The planning process is not straightforward and there are times that you will feel you are taking a step backwards. That kind of ``progress'' is normal. Additionally, there is no single right way to design a piece of research to address your research problem. Different scholars, for a variety of reasons, would end up with quite different designs for the same research problem. Design includes trade-offs, e.g., internal vs. external validity, and compromises based on time, resources, and opportunities. Knowing the subject matter – both previous research and the subject itself – helps the researcher understand where a contribution can be made and when opportunities present themselves.

Study Questions

1) Observational designs generally have higher _______ validity and lower _______ validity compared to experimental designs. Why?
2) Define spuriousness, also known as omitted variable bias.
3) Why are randomized experiments being used more and more in political science?

CHAPTER THREE: Data Collection

This chapter will introduce students to commonly used methods of data collection in political science and public policy or administration, with a particular focus on survey data. This chapter will begin with a brief discussion of quantitative vs. qualitative data collection techniques. Qualitative techniques will be described briefly, as the text primarily focuses on quantitative analysis techniques. It should be noted that data collection and analysis are two separate steps. It is possible to collect qualitative data and conduct quantitative analyses of this data. Next, the chapter will introduce some of the most frequently used types of quantitative data in political science, with as mentioned an extended discussion of survey methods.

Methods of Data Collection: Quantitative and Qualitative

Quantitative methods of data collection are those were the data are represented, often exclusively, by numbers. In stereotypical views of the field of economics, this means in numbers of dollars. One method of data collection where the data are often numbers that ultimately represent qualitative labels is survey methods. *Qualitative* methods of data collection are those where the final data product is primarily represented in words, images, or observations. These methods can often then be transformed and analyzed quantitatively such as through text analysis or coding procedures.

The divide between qualitative and quantitative data collection methods is not often clearcut, increasingly. Many researchers are relying now on what some call *mixed methods*. At its best, this means thinkign critically about how different methodologies, both qual and quant, can be used systematically to answer research questions and test hypotheses. Often, though this may mean simply research that has both a qualitative and quantitative component that are systematic in isolation but not inherently related.

Qualitative Methods of Data Collection

In political science, and the social sciences more broadly, there are a few commonly used methods of qualitative data collection that merit mentioning. This section only scratches the surface of any of these techniques. Interested readers are encouraged to seek out more authoritative texts on these topics.

The first method of qualitative data collection we will include here is *elite interviews*. Elite interviews are called as such because they focus on a population that may be hard to access. One concern unique to elite interviews is access to the population. One usually must have some kind of inside connection to interview typical elite populations such as CEOs, Congresspeople and other elected officials. Interviews can be *structured* where all questions are determined before hand. *Semi-structured* interviews are most common though where some pre-determined questions are asked and the researcher can follow interesting paths as they come up. Finally, *unstructured* interviews have very little predetermined content and are primarily exploratory and used in the early stages of projects. These data can actually be recorded, using a tape recorder and then transcribed, or may be collected through interviewer notes. *Focus groups* are similar to interviews but include group dynamics as well.

Another method of qualitative data collection is *participant observation*. This requires the researcher to sample places or contexts of interest to observe. The data are primarily collected through researcher notebooks that can either be used while observing, if it is not obtrusive or a private context, or after the fact, as soon as possible.

Document and texts are also considered qualitative data. In political science, one popular document to analyze is Congressional testimony. Other popular documents include newspapers and social media. These data are inherently qualitative. When the researcher relies primarily on their interpretations and quotes in analysis, then the analysis is qualitative as well.

For qualitative analyses of these types of data, researchers use various theory-based coding techniques that help demonstrate the patterns that emerge. This can then be quantitatively analyzed if numbers are attached to the codes (with the exception of participant observation to the authors' knowledge). This process is often done by multiple researchers with some overlapping samples to attempt to measure the agreement of the researchers on the codes present in the data. This is called *inter-rater reliability* which will be discussed again later in the text.

Quantitative Methods of Data Collection

The previous section documented a very small selection of qualitative methods of data collection with some discussion of analysis. Data that is collected in a quantitative format or method is, by defintion numeric. Sometimes those numbers have inherent meaning while other times the numbers are associated with qualitative labels.

The most obvious type of data where the numbers have inherent meaning is *financial data*. In public policy and administration, this is often budgets from governments or nonprofit organizations. In finance, researchers analyze stock prices. Economics also analyze prices and costs in various markets. Financial data, or data collected in dollar units more broadly, are convenient for analysis as you will learn later in the text. Their truly continuous nature, often to two decimal points, is valuable for many introductory methods in social science research.

Another common method of data collection that is often considered quantitative is *web scraping*. This method of data collection involves setting up a computer script (such as through R) to download a set of documents from the internet. As mentioned above, the documents themselves are qualitative. However, in web scrapping the number of documents is often so large that it is not logner considered qualitative data and requires advanced analysis techniques that are quantitative such as topic modelling or machine learning.

Finally, *surveys* are often considered a quantitative method of data collection. Surveys are similar to interviews but usually more structured and applied to a larger sample. One important distinction is the sampling which has already been discussed. Surveys, usually though not always, have larger sample sizes than interview data.

Designing Surveys

The remainder of this chapter will introduce you to some principles for designing good, scientific surveys. This again should be supplemented with further reading on the topic but will serve as a brief introduction to curious students.

When designing a survey, first the sample or target population must be determined. This decision is intertwined with research design aspects already discussed but is also important for considering the language used. For example, a survey targeting high school age children will use simpler or different language than a survey targeting a sample of the overall US population which in turn will use different language than a survey targeting university professors.

Once the survey target population has been decided, the design of the survey can begin. At this stage there are many things to consider. How is the survey being programmed or administered? Best practices for phone surveys are very different than for online surveys. What software is being used to make the survey if it is online? East Tennessee State University has access to a program called RedCap while many universities such as the University of Oklahoma us a program called Qualtrics. Other commonly used survey programming softwares, for online surveys, include SurveyMonkey and even Google

Sheets. Readers of this text are encourage to investigate these various softwares on their own. Each comes with its own sets of strengths and weaknesses that you will want to be familiar with before beginning the design of your survey.

Another important decision about survey design at this stage is length and topic. No single survey can cover all topics so you should focus your efforts on a domain of particular interest to you. This can be gender roles or environmental politics or international relations between East Asian countries and the US. A survey that attempted to address easch of these domains in depth would be too long and taxing for most respondents. Recommended lengths vary, and depend on budgets in many cases, but 20-30 minutes is generally considered a rough guideline. Time to complete can be estimated by asking friends, family, and a small sample of the relevant population to test the survey or pilot it on. These observations will not be included in the final data for analysis.

Survey Question Design

On surveys there are many types of questions you can design and use. Some are more qualitative while others are more quantitative. Some have lots of flexibility while others are relatively rigid in design. A few general principles apply to survey design.

One, for most questions, the answers should be exhaustive and mutually exclusive. More than one answer should not apply to you. If there is a question where more than one answer can apply, then respondents should be given the opportunity to say so. One way this is implemented in practice is by including a Don't Know or Not Sure option. However, this has many analytical risks as how the researcher chooses to address those respondents who choose these options can drastically affect their conclusions.

Two, scales should be consistent for similar questions when possible. This can lead to more efficient design, such as using a table in the questions, and reducing the cognitive load on respondents.

Three, extraneous text or description should be minimized whenever possible. Some questions require lengthy set-up or vignettes and therefore are exceptions to this rule. However, generally, the amount of text on a survey question should be kept to a minimum.

The list could go on and on but these three principles represent some good general rules for new survey designers.

More Specifics on Question Design

One set of relatively rigid questions is those that ask for demographics. Many researchers, such as those who collected the data used in this text, attempt to use the questions asked by the U.S. Census Bureau. For example, the way race is asked on the data used in this text is similar to the Census questions. In particular, the separation of Hispanic as an ethincity separate from the race question follows these norms. Demographic questions also provide a good venue to bring up survey ordering. Questions at the beginning are most likely to be finished. Thus, you generally want to put questions of highest importance early in the survey. Demographics are tricky in this regard. They are often vital for social science research but putting them at the beginning of the survey may lead to non-response on other more substantive questions. In political science, the placement of ideology and partisanship questions are also vital. They are key to many research questions but putting them early in the survey may both some respondents and cause them to stop responding as well. In the data for this project, most demographic questions were asked in the first pages while political questions were asked in the last.

Among demographic questions we see some variety. Questions like the race question used in this text are what can be called *closed ended questions with an open ended response* as most options are stated by the survey. There is, however, one opended ended option that is Other which requires respondents to type in their race that is not in the list. A purely *open ended* demographic question is income or age. In both cases, respondents must type in (or on a phone survey, state) verbatim their age and income. Other open ended questions give respondents a text box, if online, or just time to state their answer. These questions result in data that is qualitative. Most other questions result in quantitative data because the qualitative label, say African American for race, is translated to a number that represents that label, say 2. These numbers can then be used in quantitative analyses.

Choosing between close and open ended questions requires researchers to prioritize their research questions and desired data. Open ended questions result in more nuanced, deeper data but cannot be analyzed, as easily, with typical quantiative methods such as those taught in this text.

These sections only scratch the surface of survey design. Other concerns include how the answers are formatted (radio button, slider, etc.), appropriate length of scales (5? 7? 10?), experimental designs (how many treatments?) and many more. Hopefully, this chapter will help students better understand about the complexities of data collection, either in their own project or for the data used in this text. So much of the work occurs before the data is ever even collected.

Study Questions

1) Design a survey question that is close-ended. Be sure to apply the principles of design and other recommendations from this chapter.
2) Design a survey question that is open-ended. Be sure to apply the principles of design and other recommendations from this chapter.
3) What qualitative method is most difficult to analyze quantitatively? Why?

CHATPER FOUR: Downloading and Getting Started with Excel for Statistics

This chapter will introduce you to the basics of using Excel in the textbook. In particular, this chapter will demonstrate how to access necessary add-ons.

Introduction to Excel

Excel is a Microsoft program used for many purposes though primarily for spreadsheet management. It can be use for budgeting, data management, and even statistical analysis – the topic of this book. Excel is a great tool for a number of reasons:

- commonly found on most computers and provided by most workplaces,
- great for visually examining data, and
- graphical facilities for data analysis and display either on-screen or on hardcopy

Excel is not as powerful or full featured as many of its competitors. However, its prevalence makes knowledge of it really valuable in the modern workplace.

Downloading ToolPak

Despite Excel's many benefits, it does require an additional download or package to take full advantage of its statistical capabilities. In this section we will provide instructions to downloading this add-on called ToolPak.

From Standalone, Desktop Based Excel

First, you will need to cick on the File tab on the left side of the toolbar at the top of the screen. Then, click on Options at the bottom of the list to the left of the screen. On the left sidebar within the dialog box, you will choose Add-ins. After this, in the center section, there should be a list labelled Inactive Application Add-ins. Toward the top of this list you should see and click on "Analysis ToolPak – VBA" (There is also an option for "Analysis Toolpak". If you only have one option, either should be fine). Then at the bottom of the dialog box, next to the drop-down menu labelled Manage, click Go… And then check the box next to "Analysis ToolPak – VBA" (or "Analysis ToolPak"). At this point, select OK.

At this point, the ToolPak can now be found under the "Data" tab of the toolbar across the top of your screen. It will be located on the far right end and will be labelled "Data Analysis". Clicking on the Data Analysis link will bring up a dialog box containing a list of functions the ToolPak can perform.

Downloading the Data Analysis Toolpak from Office 365

First, click on the "Insert" tab in the middle of the toolbar at the top of the screen. Then, select "Office Add-ins". In the dialog box that appears, select the "STORE" tab (located at the top right). In the search bar on the left, type "analysis toolpak". From the results, find "XLMiner Analysis ToolPak" in the list in the center of the dialog box (it should be the only program that comes up). Click "Add". The ToolPak should appear as a sidebar on the right side of your screen with a list of options similar to the Analysis ToolPak in the standalone version of Excel.

To access the ToolPak later from another device, you will need to repeat the first steps of this process, then go to the "MY ADD-INS" tab of the Office Add-ins dialog box.For the additional device, select XLMiner Analysis ToolPak (likely the only program listed) and then click “Add” at the bottom of the dialog box.

Data in Excel

Excel can handle a few different file types as data. The primary type that will be used for the book and accompanying course is a comma separated file, or .csv file type. A CSV is a convenient file type that is portable across many operating platforms (Mac, Windows, etc) as well as statistical/data manipulation softwares. Other common file types are text (.txt) and Excel files (.xls or .xlsx). Each of these can be opened easily in Excel. Some more advanced statistical softwares require their own data file type. These can often, with some care, be opened in Excel as well.

For the purposes of the book, we will acquire our data by going here. You will then type your e-mail where it says Request Data. You should then receive an e-mail with the data attached as a .csv file. First, you will want to download this data onto your computer. We recommend creating a folder specifically for the book and its data (and if you’re in the class for your classwork).

Data in Manipulation in Excel

Excel is a very flexible tool for manipulating data into various subsets and forms. Excel will allow users to transform their data from long to wide formats, remove NA values, recode variables, etc. In order to make the downloaded data more manageable for the book, we are going to do two things. First, we want to restrict our data to one wave. The data we downloaded represent many waves of a quarterly survey that is sent to a panel of Oklahoma residents on weather, climate and policy preferences. This book will not venture into panel data analysis or time series analysis, as it is an introductory text, and therefore we simply want one cross section of data for our analysis.

To do this, go to the Data Tab and choose Filter. Unselect the option for Wave 12 (Fall 2016). Then highlight the first row by clicking on the number to the left of the first column. Then, hit control-shift-down arrow to highlight all remaining rows. Then press F5 (or the Mac equivalent). This will open the Go To box. Click Special in the bottom left. Then choose "Visible cells only" and click okay. Then, right click on the selection and choose delete row. Now remove the filter and you should be left with only the observations of one wave of the original panel data. If this doesn't work, the Class Data Set will also be provided to students in the course and can be access from the original github site for the original text.

Saving and Writing Data

Saving or writing data that we have manipulated is a useful tool. It allows us to easily share datasets we have created with others. This is useful for collaboration. Additionally, this will be useful for the book, as our new dataset is the one that will be worked with throughout the book. This dataset is much smaller than the one we originally downloaded and therefore will allow for quicker load times as well as hopefully reduce potential confusion.

Students using Excel should save various versions of their data if they make changes to the data's structure. These files should have names that are descriptive for the student or researcher and include the date they were saved on. This will allow the researcher student to go back and examine the changes they made over time. It is best to generally add on new variables or data without getting rid of old dat if possible, with the exception of this initial subset. This means you won't ever lose data you might need again someday but don't think you need now.

Study Questions

1) Do you have the ToolPak downloaded on your personal computer (laptop or desktop)? If not, why not?

2) Why is Excel a useful software to learn?

CHAPTER FIVE: Exploring and Visualizing Data

You have your plan, you carry out your plan by getting out and collecting your data, and then you put your data into a file. You are excited to test your hypothesis, so you immediately run your multiple regression analysis and look at your output. You can do that (and probably will even if we advise against it), but before you can start to make sense of that output you need to look carefully at your data. You will want to know things like "how much spread do I have in my data" and "do I have any outliers". (If you have limited spread, you may discover that it is hard to explain variation in something that is nearly a constant and if you have an outlier, your statistics may be focused on trying to explain that one case.)

In this chapter, we will identify the ways to characterize your data before you do serious analysis, both to understand what you are doing statistically and to error-check.

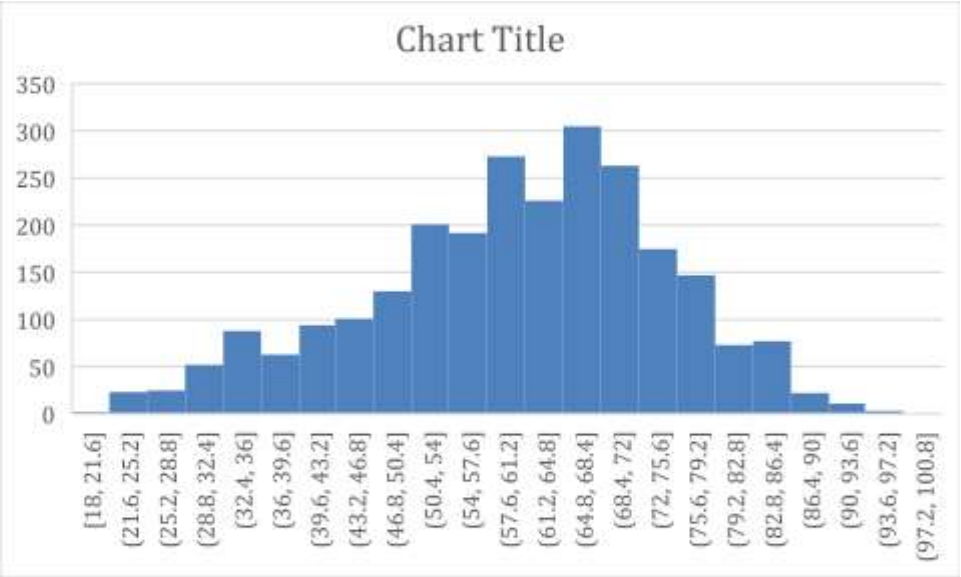
Characterizing Data

What does it mean to characterize your data? First, it means knowing **how many observations** are contained in your data and **the distribution** of those observations over the range of your variable(s). What kinds of measures (interval, ordinal, nominal) do you have, and what are the ranges of valid measures for each variable? How many cases of missing (no data) or mis-coded (measures that fall outside the valid range) do you have? What do the coded values represent? While seemingly trivial, checking and evaluating your data for these attributes can save you major headaches later. For example, missing values for an observation often get a special code -- say, "-99" -- to distinguish them from valid observations. If you neglect to treat these values properly, Excel (or any other statistics program) will treat that value as if it were valid and thereby turn your results into a royal hairball. We know of cases in which even seasoned quantitative scholars have made the embarrassing mistake of failing to properly handle missing values in their analyses. In at least one case, a published paper had to be retracted for this reason. So don't skimp on the most basic forms of data characterization!

The dataset used for purposes of illustration in this version of this text is taken from a survey of Oklahomans, conducted in 2016, by the University of Oklahoma's Center for Risk and Crisis Management. The survey question wording and background will be provided in class. However, for purposes of this chapter, note that the measure of `ideology` consists of a self-report of political ideology on a scale that ranges from 1 (strongly liberal) to 7 (strongly conservative); the measure of the `perceived risk of climate change` ranges from zero (no risk) to 10 (extreme risk). `Age` was measured in years.

It is often useful to graph the variables in your dataset to get a better idea of their distribution. In addition, we may want to compare the distribution of a variable to a theoretical distribution (typically a normal distribution). This can be accomplished in several ways, but we will show two here---a histogram and a density curve---and more will be discussed in later chapters. For now we examine the distribution of the variable measuring age. The red line on the density visualization presents the normal distribution given the mean and standard deviation of our variable.

A histogram creates intervals of equal length, called bins, and displays the frequency of observations in each of the bins. To produce a histogram in Excel first go to the Insert tab and then charts in Excel. For a histogram, you will choose the Insert Statistic Chart which is the middle icon of the small icons. Then choose the first option histogram. This will result in a very bare bones chart such as the one below.



A plot such as this should be polished before being shown to any important end audience. Both axes should be labeled appropriately. In this case, the X-axis (horizontal) should be labelled age and the y-axis (vertical) should be labelled frequency. These can be added using the Axis Titles menu which can be accessed by clicking on the big Plus sign to the right of the figure and choosing Axis Titles. Text boxes will populate on the figure and you can type in appropriate titles. The Chart Title should also be replaced with a general main title that is informative such as "Histogram of Age for Survey Respondents". In some cases, this will not be necessary as you will be told to put the title in text below the figure. This figure can be transformed into a density function with some effort in Excel. A density plot is similar but instead of bars it plots a line and the y-axis is probability density instead of frequency.

You can also get an overview of your data using a table known as a frequency distribution. The frequency distribution summarizes how often each value of your variable occurs in the dataset. If your variable has a limited number of values that it can take on, you can report all values, but if it has a large number of possible values (e.g., age of respondent), then you will want to create categories, or bins, to report those frequencies. In such cases, it is generally easier to make sense of the percentage distribution. The table below is a frequency distribution for the ideology variable. From that table we see, for example, that about one-third of all respondents are moderates. We see the numbers decrease as we move away from that category, but not uniformly. There are a few more people on the conservative extreme than on the liberal side and that the number of people placing themselves in the penultimate categories on either end is greater than those towards the middle. The histogram and density curve would, of course, show the same pattern.

The other thing to watch for here (or in the charts) is whether there is an unusual observation. If one person scored 17 in this table, you could be pretty sure a coding error was made somewhere. You cannot find all your errors this way, but you can find some, including the ones that have the potential to most seriously adversely affect your analysis.

A frequency table can be made using the Pivot Table function in Excel. First, select all of our data. This can be done by highlighting the first column by clicking on it then using ctrl+shift+down and then ctrl+shift+right. Then go to the Insert Tab. Then choose Pivot Table. This should pop up the Pivot Table commands on the right of the screen. From here, search for your preferred variable, in this case ideol. Drag and drop this to the rows field and the values field. Drag and drop ideol to the values field two more times. On the second one, click and choose Value Field Settings. Then click on Show Values As. Then from the dropdown menu choose % of Column Total. Repeat these steps with the third ideology and instead choose % of Running Total. This will return the following table.

Row Labels	Count of ideol	Count of ideol2	Count of ideol3
1	122	4.79%	4.79%
2	279	10.95%	15.74%
3	185	7.26%	23.01%
4	571	22.42%	45.43%
5	328	12.88%	58.30%
6	688	27.01%	85.32%
7	351	13.78%	99.10%
NA	23	0.90%	100.00%

(blank)		0.00%	100.00%
Grand Total	**2547**	**100.00%**	

As above, this table should be polished by removing the blank row and making nicer, more descriptive column titles. Having obtained a sample, and described the frequency of key variables, it is important to be able to characterize that sample other ways. In particular, it is important to understand the probability distributions associated with each variable in the sample.

Central Tendency

Measures of central tendency are useful because a single statistic can be used to describe the distribution. We focus on three measures of central tendency: the mean, the median, and the mode.

Measures of Central Tendency

The Mean: The arithmetic average of the values

The Median: The value at the center of the distribution

The Mode: The most frequently occurring value

We will primarily rely on the mean, because of its efficient property of representing the data. But medians – particularly when used in conjunction with the mean - can tell us a great deal about the shape of the distribution of our data. We will return to this point shortly.

Level of Measurement and Central Tendency

The three measures of central tendency – the mean, median, and mode – each tell us something different about our data, but each has some limitations as well (especially when used alone). Knowing the mode tells us what is most common, but we do not know how common and, using it alone, would not even leave us confident that it is an indicator of anything very *central*. When rolling in your data, it is generally a good idea to roll in all the descriptive statistics that you can to get a good feel for them.

One issue, though, is that your ability to use any statistic is dependent on the level of measurement for the variable. The mean requires you to add all your observations together. But you cannot perform mathematical functions on ordinal or nominal level measures. Your data must be measured at the interval level to calculate a meaningful mean. (If you ask Excel to calculate the mean student id number, it will, but what you get will be nonsense.) Finding the middle item in an order listing of your observations (the median) requires the ability to order your data, so your level of measurement must be at least ordinal. Therefore, if you have nominal level data, you can only report the mode (but no median or mean), so it is critical that you also look beyond central tendency to the overall distribution of the data.

Moments

In addition to measures of central tendency, “moments” are important ways to characterize the shape of the distribution of a sample variable. Moments are applicable when the data measured is interval type (the level of measurement). The first four moments are those that are used most often.

The First Four Moments

1. *Expected Value:* The expected value of a variable, *E(X)* is its mean.

$$E(X) = \bar{x} = \frac{\Sigma X_i}{n}$$

2. *Variance:* The variance of a variable concerns the way that the observed values are spread around either side of the mean.

$$s_x^2 = \frac{\Sigma(X - \bar{X})^2}{n-1}$$

3. *Skewness:* The skewness of a variables is a measure of its asymmetry.

$$S = \frac{\Sigma(X - \bar{X})^3}{n-1}$$

4. *Kurtosis:* The kurtosis of a variable is a measure of its peakedness.

$$K = \frac{\Sigma(X - \bar{X})^4}{n-1}$$

The First Moment – Expected Value

The *expected value* of a variable is the value you would obtain if you could multiply all possible values within a population by their probability of occurrence. Alternatively, it can be understood as the mean value for a population variable. An expected value is a theoretical number, because we usually cannot observe all possible occurrences of a variable. The mean value for a sample is the average value for the variable X, and is calculated by adding the values of X and dividing by the sample size n. In Excel, this can be achieved by typing into an empty cell =AVERAGE(range of data) and highlighting the variable for which you would like to calculate a mean.

The Second Moment – Variance and Standard Deviation

The *variance* of a variable is a measure that illustrates how a variable is spread, or distributed, around its mean. The population variance is expressed as σ_x^2while the sample variance is presented the same but replacing the lower-case sigma with a lowercase s.

Variance is measured in *squared* deviations from the mean, and the sum of these squared variations is termed the **total sum of squares**. Why squared deviations? Why not just sum the differences? While the latter strategy would seemingly be simpler, it would always sum to zero. By squaring the deviations we make them all positive, so the sum of squares will always be a positive number.

Total Sum of Squares is the squared summed total of the variation of a variable around its mean.

This can be expressed as:

$$TSS_x = \sum(X_i - \overline{X})^2$$

Therefore,

$$s_x^2 = \frac{TSS_x}{n-1}$$

The square root of the variance is called the standard deviationσ_x . The sample s.d. is expressed as:

$$s_x = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}.$$

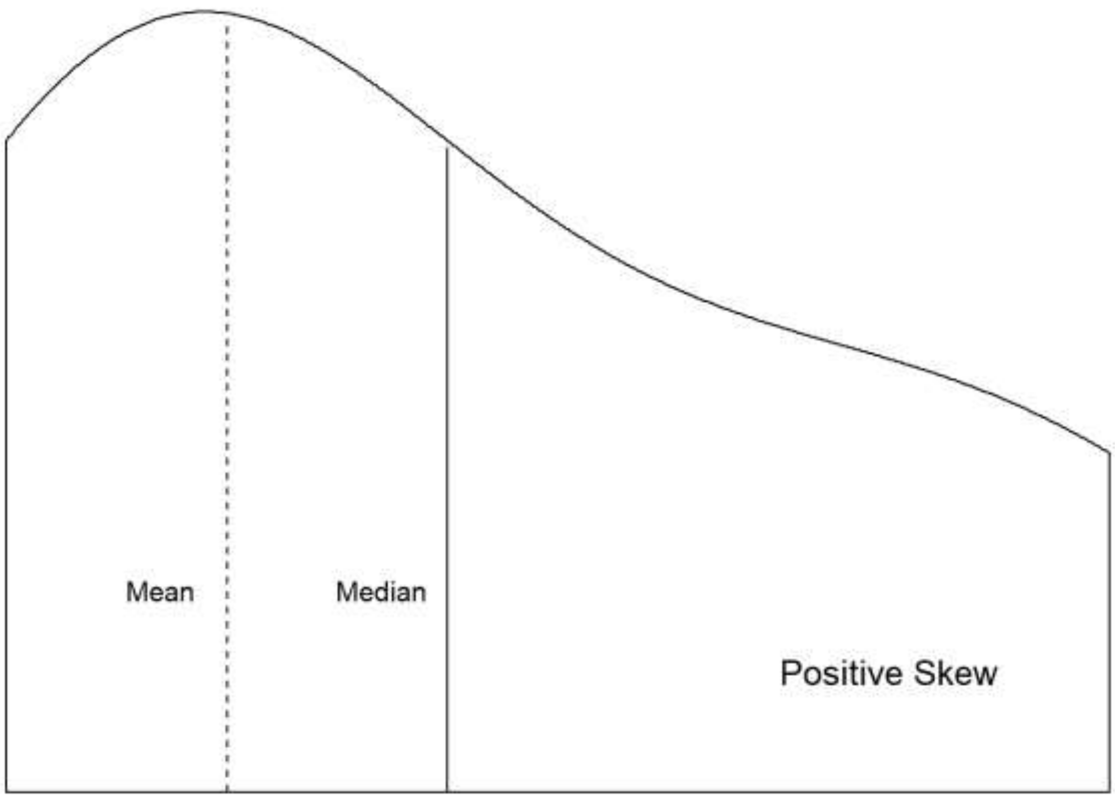
This can be calculated in Excel using =VAR.S(Range of the Variable) and = STDEV.S(Range of Variable). The .S versions are used because we are working with a sample and not a population.

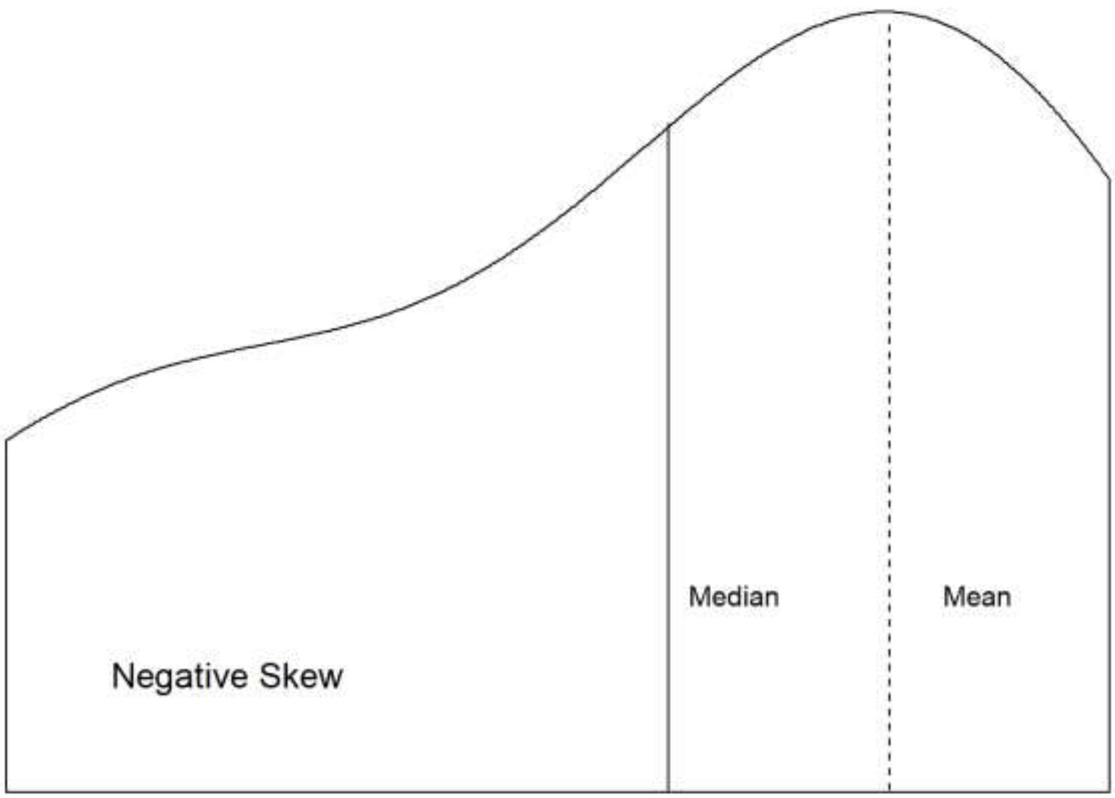
The Third Moment – Skewness

Skewness is a measure of the asymmetry of a distribution. Specifically, skewness refers to the position of the expected value (i.e., mean) of a variable distribution relative to its median. It is calculated as:

$$S = \frac{\frac{\Sigma(X - \bar{X})^3}{n-1}}{\left(\sqrt{\frac{\Sigma(X-\bar{X})^2}{(n-1)}}\right)^3}$$

When the mean and median of a variable are roughly equal, then the Mean ~~ Median; the distribution is approximately symmetrical and S = 0. This means an equal proportion of the variable lies on either side of the mean. However, if the Mean > Median then the variable has a positive skew and S > 0. If the Median > Mean the then S < 0 and the variable has a negative skew.


Mean
Median
Positive Skew



The Fourth Moment – Kurtosis

The *kurtosis* of a distribution refers to the the peak of a variable (i.e., the mode) and the relative frequency of observations in the tails. It is calculated:

$$K = \frac{\frac{\Sigma(X - \bar{X})^4}{(n-1)}}{\left(\frac{\Sigma(X - \overline{X})^2}{(n-1)}\right)^2}$$

In general, higher kurtosis is indicative of a distribution where the variance is a result of low frequency yet more extreme observed values. In addition, when K<3, the distribution is *platykurtic,* which is flatter and/or more “short-tailed” than a normal distribution. When K > 3, the distribution is *leptokurtic,* which is a slim, high-peak and long tails. For a normal distribution, K = 3.

Order Statistics

Apart from central tendency and moments, probability distributions can also be characterized by **order statistics**. Order statistics are based on the position of a value in an ordered list. Typically, the list is ordered from low values to high values.

Order Statistics

Summaries of values based on position in an ordered list of all values. Types of order statistics include the minimum value, the maximum value, the median, quartiles, and percentiles.

- *Minimum Value*: The lowest value of a distribution
- *Maximum Value*: The highest value of a distribution
- *Median*: The value at the center of a distribution
- *Quartiles*: Divides the values into quarters
- *Percentiles*: Divides the values into hundredths

Median

The *median* is the value at the center of the distribution, therefore 50% of the observations in the distribution will have values above the median and 50% will have values below. For samples with a n-size that is an odd number, the median is simply the value in the middle. For example, with a sample consisting of the observed values of 1,2,3,4,5. the median is 3. Distributions with an even numbered n-size, the median is the average of the two middle values. The median of a sample consisting of the observed values of 1,2,3,4,5,6 would be (3+4)/2 or 3.5.

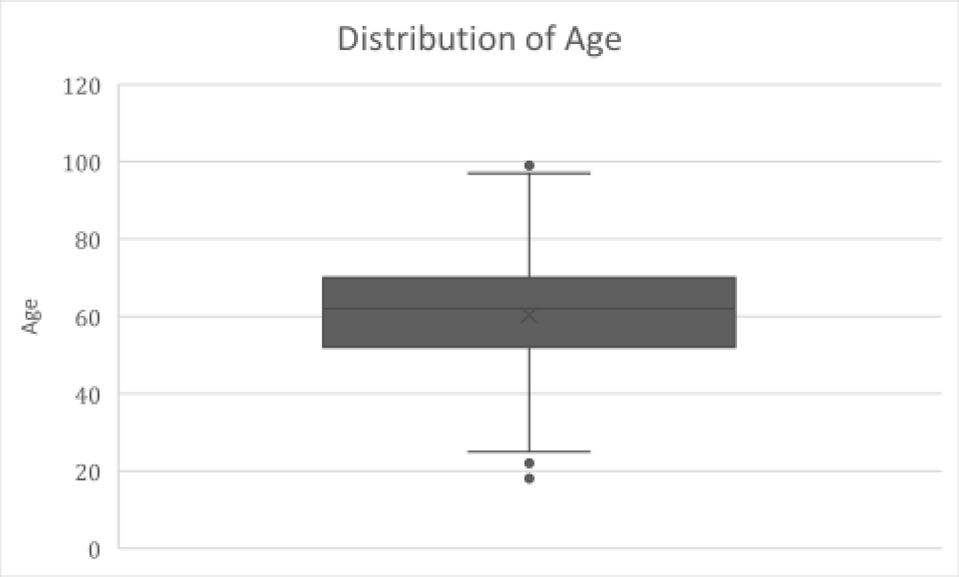
The median is the order statistic for central tendency. In addition, it is more “robust” in terms of extreme values than the mean. Extremely high values in a distribution can pull the mean higher, and extremely low values pull the mean lower. The median is less sensitive to these extreme values. The median is therefore the basis for “robust estimators”, to be discussed later in this book.

Quartiles

Quartiles split the observations in a distribution into quarters. The first quartile, Q1, consists of observations whose values are within the first 25% of the distribution. The values of the second quartile, Q2, are contained within the first half (50%) of the distribution, and is marked by the distribution’s median. The third quartile, Q3, includes the first 75% of the observations in the distribution.

The interquartile range (IQR) measures the spread of the ordered values. It is calculated by subtracting Q1from Q3, or IQR = Q3-Q1.

We can visually examine the order statistics of a variable with a boxplot. A boxplot displays the range of the data, the first and third quartile, the median, and any outliers. This can be done in Excel using the Insert tab, Charts, Statistic Chart, Box and Whisker Chart. For our data, for age, the chart below is created with a few modifications including using gray-scale (this color scheme is greatly preferred to the default).



Percentiles

Percentiles- list the data in hundredths. For example, scoring in the 99th percentile on the GRE means that 99% of the other test takers had a lower score. Percentiles can be incorporated with quartiles (and/or other order statistics) such that: - First Quartile: 25th percentile - Second Quartile: 50th percentile (the median) - Third Quartile: 75th percentile. These can be found in Excel by choosing an open cell and typing in:

= PERCENTILE.EXC(range of data,0.25)

= PERCENTILE.EXC(range of data,0.5)

= PERCENTILE.EXC(range of data,0.75).

The second part of this function can range anywhere between 0 and 1 to acquire any percentile, not just quartiles. We can also find the minimum and maximum of the data or variable using similar functions:

=MIN(Range of data)

=MAX(Range of data).

Summary

It is a serious mistake to begin your data analysis without understanding the basics of your data. Knowing their range, the general distribution of your data, the shape of that distribution, their central tendency, and so forth will give you important clues as you move through your analysis and interpretation and prevent serious errors from occurring. Readers also often need to know this information to provide a critical review of your work.

Overall, this chapter has focused on understanding and characterizing data. We refer to the early process of evaluating a data set as rolling in the data – getting to know the characteristic shapes of the distributions of each of the variables, the meanings of the scales, and the quality of the observations. The discussion of central tendency, moments, and order statistics are all tools that you can use for that purpose. As a practicing scholar, policy analyst, or public administration practitioner, this early stage in quantitative analysis is not optional; a failure to carefully and thoroughly understand your data can result in analytical disaster, excruciating embarrassment, and maybe even horrible encounters with the Killer Rabbit of Caerbannog.

Think of rolling in the data, then, as your version of the Holy Hand Grenade of Antioch.

Study Questions

1. Define the mean using both mathematical notation and words.
2. What measures of central tendency can be applied to continuous (interval and ratio) data? Which measures of central tendency can be applied to ordinal data? Which measures of central tendency can be applied to nominal/categorical data?
3. Why is digging into the data and the distribution of your data an important first (or early) step in your analysis?
4. What are the third and fourth moments of a distribution? What do they tell us?

CHAPTER SIX: Inference

This chapter considers the role of inference—learning about populations from samples—and the practical and theoretical importance of understanding the characteristics of your data before attempting to undertake statistical analysis. As we noted in the prior chapters, it is a vital first step in empirical analysis to “roll in the data.”

Inference: Populations and Samples

The basis of hypothesis testing with statistical analysis is **inference**. In short, inference—and inferential statistics by extension—means deriving knowledge about a population from a sample of that population. Given that in most contexts it is not possible to have all the data on an entire population of interest, we therefore need to sample from that population.[1] However, in order to be able to rely on inference, the sample must cover the theoretically relevant variables, variable ranges, and contexts.

Populations and Samples

In doing statistical analysis we differentiate between populations and samples. The population is the total set of items that we care about. The sample is a subset of those items that we study in order to understand the population. While we are interested in the population we often need to resort to studying a sample due to time, financial, or logistic constraints that might make studying the entire population infeasible. Instead, we use inferential statistics to make inferences about the population from a sample.

Sampling and Knowing

Take a relatively common – but perhaps less commonly examined – expression about what we “know” about the world around us. We commonly say we “know" people, and some we know better than others. What does it mean to know someone? In part it must mean that we can anticipate how that person would behave in a wide array of situations. If we know that person from experience, then it must be that we have observed their behavior across a sufficient variety of situations in the past to be able to infer how they would behave in future situations. Put differently, we have “sampled” their behavior across a relevant range of situations and contexts to be confident that we can anticipate their behavior in the future.[2] Similar considerations about sampling might apply to “knowing” a place, a group, or an institution. Of equal importance, samples of observations across different combinations of variables are necessary to identify relationships (or functions) between variables. In short, samples – whether deliberately drawn and systematic or otherwise – are integral to what we think we know of the world around us.

Sampling Strategies

Given the importance of sampling, it should come as little surprise that there are numerous strategies designed to provide useful inference about populations. For example, how can we judge whether the temperature of a soup is appropriate before serving it? We might stir the pot, to assure uniformity of temperature across possible (spoon-sized) samples, then sample a spoonful. A particularly thorny problem in sampling concerns the practice of courtship, in which participants may attempt to put “their best foot forward” to make a good impression. Put differently, the participants often seek to bias the sample of relational experiences to make themselves look better than they might on average. Sampling in this context usually involves (a) getting opinions of others, thereby broadening (if only indirectly) the size of the sample, and (b) observing the courtship partner over a wide range of circumstances in which the intended bias may be difficult to maintain. Put formally, we may try to stratify the sample by taking observations in appropriate “cells” that correspond to different potential influences on behavior – say, high stress environments involving preparation for final exams or meeting parents. In the best possible case, however, we try to wash out the effect of various influences on our samples through randomization. To pursue the courtship example (perhaps a bit too far!), observations of behavior could be taken across interactions from a randomly assigned array of partners and situations. But, of course, by then all bets are off on things working out anyway.

Sampling Techniques

When engaging in inferential statistics to infer about the characteristics of a population from a sample, it is essential to be clear about how the sample was drawn. Sampling can be a very complex practice with multiple stages involved in drawing the final sample. It is desirable that the sample is some form of a **probability sample**, i.e., a sample in which each member of the population has a known probability of being sampled. The most direct form of an appropriate probability sample is a **random sample** where everyone has the same probability of being sampled. A random sample has the advantages of simplicity (in theory) and ease of inference as no adjustments to the data are needed. But, the reality of conducting a random sample may make the process quite challenging. Before we can draw subjects at random, we need a list of all members of the population. For many populations (e.g. adult US residents) that list is impossible to get. Not too long ago, it was reasonable to conclude that a list of telephone numbers was a reasonable approximation of such a listing for American households. During the era that landlines were ubiquitous, pollsters could randomly call numbers (and perhaps ask for the adult in the household who had the most recent birthday) to get a good approximation of a national random sample. (It was also an era before caller identification and specialized ringtones, which meant that calls were routinely answered, therefore decreasing - but not eliminating - concern with response bias.) Of course, telephone habits have changed and pollsters find it increasingly difficult to make the case that random dialing of landlines serves as a representative sample of adult Americans.

Other forms of probability sampling are frequently used to overcome some of the difficulties that pure random sampling presents. Suppose our analysis will call upon us to make comparisons based on race. Only 12.6% of Americans are African-American. Suppose we also want to take into account religious preference. Only 5% of African-Americans are Catholic, which means that only .6% of the population is both. If our sample size is 500, we might end up with three Catholic African-Americans. A **stratified random sample** (also called a quota sample) can address that problem. A stratified random sample is similar to a simple random sample, but will draw from different subpopulations, strata, at different rates. The total sample needs to be weighted, then, to be representative of the entire population.

Another type of probability sample that is common in face-to-face surveys relies on **cluster sampling**. Cluster sampling initially samples based on clusters (generally geographic units, such as census tracts) and then samples participants within those units. In fact, this approach often uses multi-level sampling where the first level might be a sample of congressional districts, then census tracts, and then households. The final sample will need to be weighted in a complex way to reflect varying probabilities that individuals will be included in the sample.

Non-probability samples, or those for which the probability of inclusion of a member of the population in the sample is unknown, can raise difficult issues for statistical inference; however, under some conditions, they can be considered representative and used for inferential statistics.

Convenience samples (e.g., undergraduate students in the Psychology Department subject pool) are accessible and relatively low cost, but may differ from the larger population to which you want to infer in important respects. Necessity may push a researcher to use a convenience sample, but inference should be approached with caution. A convenience sample based on “I asked people who came out of the bank” might provide quite different results from a sample based on “I asked people who came out of a payday loan establishment”.

Some non-probability samples are used because the researcher does not want to make inferences to a larger population. A **purposive or judgmental sample** relies on the researcher's discretion regarding who can bring useful information to bear on the subject matter. If we want to know why a piece of legislation was enacted, it makes sense to sample the author and co-authors of the bill, committee members, leadership, etc., rather than a random sample of members of the legislative body.

Snowball sampling is similar to a purposive sample in that we look for people with certain characteristics but rely on subjects to recommend others who meet the criteria we have in place. We might want to know about struggling young artists. They may be hard to find, though, since their works are not hanging in galleries so we may start with a one or more that we can find and then ask them who else we should interview.

Increasingly, various kinds of non-probability samples are employed in social science research, and when this is done it is critical that the potential biases associated with the samples be evaluated. But there is also growing evidence that non-probability samples can be used inferentially - when done very carefully, using complex adjustments. Wang, et al. (2014) demonstrate that a sample of Xbox users could be used to forecast the 2012 presidential election outcome. [3] An overview of their technique is relatively simple, but the execution is more challenging. They divided their data into cells based on politically and demographically relevant variables (e.g., party id, gender, race, etc.) and ended up with over 175,000 cells - poststratification. (There were about three-quarters of a million participants in the Xbox survey). Basically, they found the vote intention within each cell and then weighted each cell based on a national survey using multilevel regression. Their final results were strikingly accurate. Similarly, Nate Silver, with FiveThirtyEight, has demonstrated remarkable ability to forecast based on his weighted sample of polls taken by others.

Sampling techniques can be relatively straightforward, but as one moves away from simple random sampling, the sampling process either becomes more complex or limits our ability to draw inferences about a population. Researchers use all of these techniques for good purposes and the best technique will depend on a variety of factors, such as budget, expertise, need for precision, and what research question is being addressed. For the remainder of this text, though, when we talk about drawing inferences, the data will be based upon an appropriately drawn probability sample.

So How is it That We Know?

So why is it that the characteristics of samples can tell us a lot about the characteristics of populations? If samples are properly drawn, the observations taken will provide a range of values on the measures of interest that reflect those of the larger population. The connection is that we expect the phenomenon we are measuring will have a **distribution** within the population, and a sample of observations drawn from the population will provide useful information about that distribution. The theoretical connection comes from probability theory, which concerns the analysis of random phenomena. For present purposes, if we randomly draw a sample of observations on a measure for an individual (say, discrete acts of kindness), we can use probability theory to make inferences about the characteristics of the overall population of the phenomenon in question. More specifically, probability theory allows us to make inference about the shape of that distribution – how frequent are acts of kindness committed, or what proportion of acts evidence kindness?

In sum, samples provide information about **probability distributions**. Probability distributions include all possible values and the probabilities associated with those values. The **normal distribution** is the key probability distribution in inferential statistics.

The Normal Distribution

For purposes of statistical inference, the normal distribution is one of the most important types of probability distributions. It forms the basis of many of the assumptions needed to do quantitative data analysis, and is the basis for a wide range of hypothesis tests. A standardized normal distribution has a mean, μ, of 0 and a standard deviation (s.d.), σ, of 1. The distribution of an outcome variable, Y, can be described:

$$Y \sim N(\mu_y, \sigma^2)$$

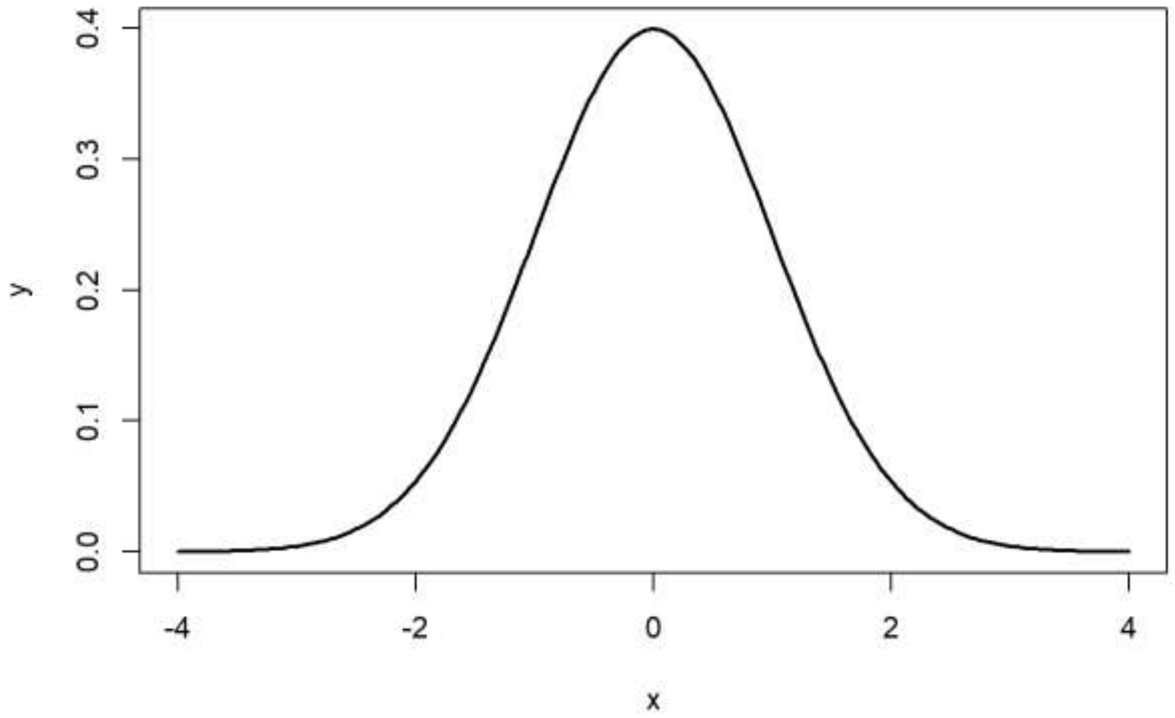
Where ~ stands for "distributed as", N indicates the normal distribution, and the mean μ_y and variance σ^2 are the parameters. The probability function of the normal distribution is expressed below:

The Normal Probability Density Function: The probability density function (PDF) of a normal distribution with mean μ and standard deviation σ:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-.5\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

The Standard Normal Probability Density Function: The standard normal PDF has a μ=0 and σ=1 is represented in equation and graphical form below:

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-.5(x\,)^2}$$



Note that the tails go to ±∞±∞. In addition, the density of a distribution over the range of x is the key to hypothesis testing. With a normal distribution, ∼68%∼ of the observations will fall within 1 standard deviation of the mean, ∼95% will fall within 2 standard deviations, and ∼99.7%within 3 standard deviations. This is illustrated below:

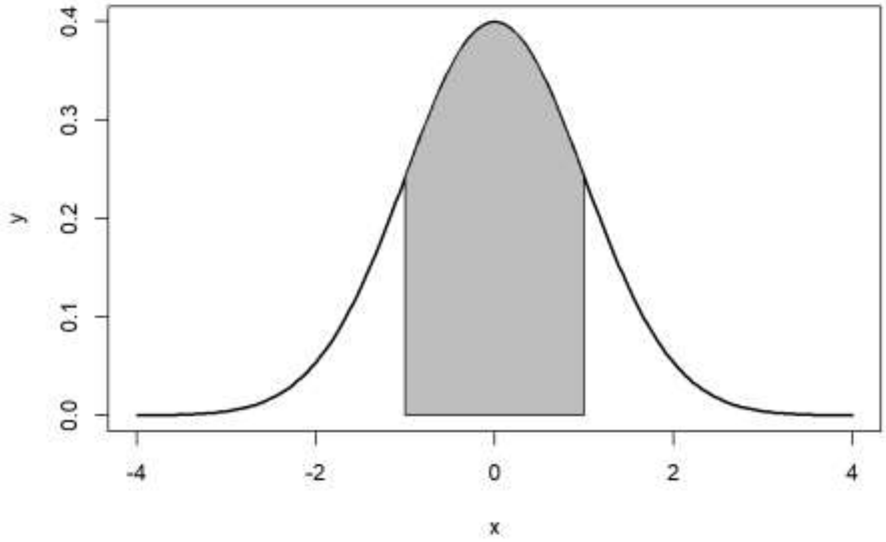


Figure ~68%, 1 standard deviation

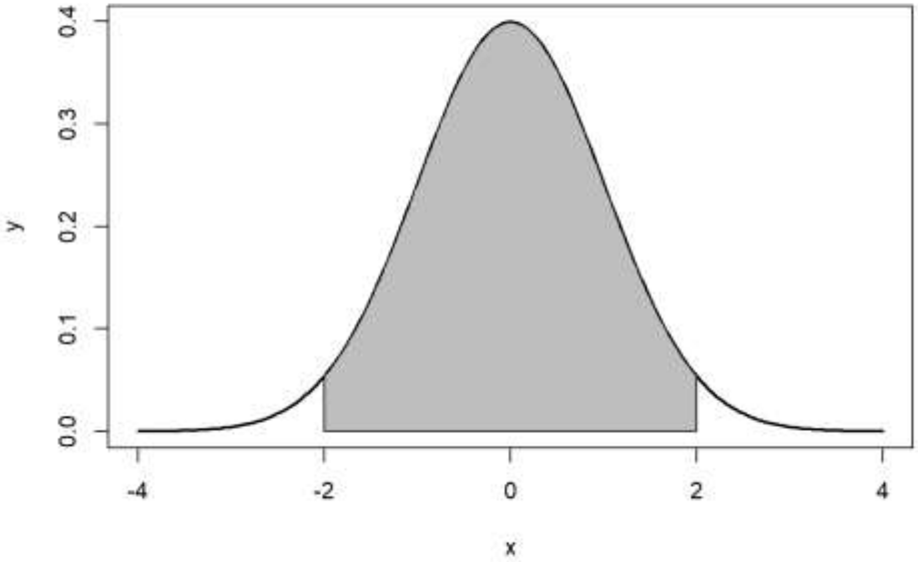
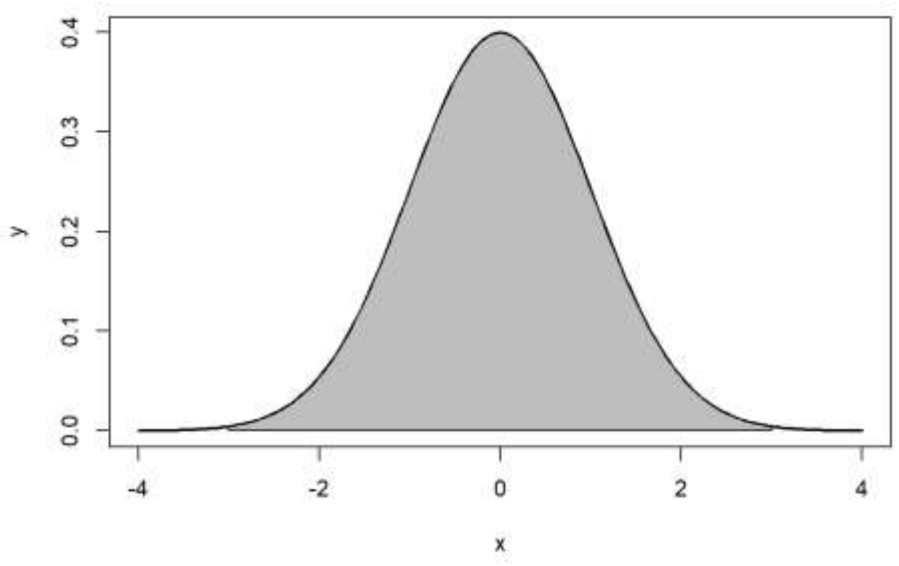


Figure: 95%, 2 standard deviations



Figure, 99.7%, 3 standard deviations

The normal distribution is characterized by several important properties. The distribution of observations is symmetrical around the mean μ; the frequency of observations is highest (the mode) at μ, with more extreme values occurring with lower frequency; and only the mean and variance are needed to characterize data and test simple hypotheses.

The Properties of the Normal Distribution

- It is symmetrical around its mean and median, μ
- The highest probability (aka "the mode") occurs at its mean value
- Extreme values occur in the tails
- It is fully described by its two parameters, μ and σ^2

If the values for μ and $\sigma^2$are known, which might be the case with a population, then we can calculate a Z-score to compare differences in μ and σ^2 between two normal distributions or obtain the probability for a given value given μ and σ^2. The Z-score is calculated:

$$Z = \frac{Y - \mu_y}{\sigma}$$

Therefore, if we have a normal distribution with a μ of 70 and a σ^2 of 9, we can calculate a probability for i=75. First, we calculate the Z-score, then we determine the probability of that score based on the normal distribution. In Excel, this can be done using the simple math commands at first. To find the probability or p-value of this in Excel, type into an empty cell, =NORM.DIST(75,70,3,TRUE). This returns a probability of 0.95221 which means observation is just outside of 2 standard deviations away. To calculate the sorts of p-values used later in the book, simply calculate 1-0.95221 = 0.04779 which is less than the cut-off of 0.05.

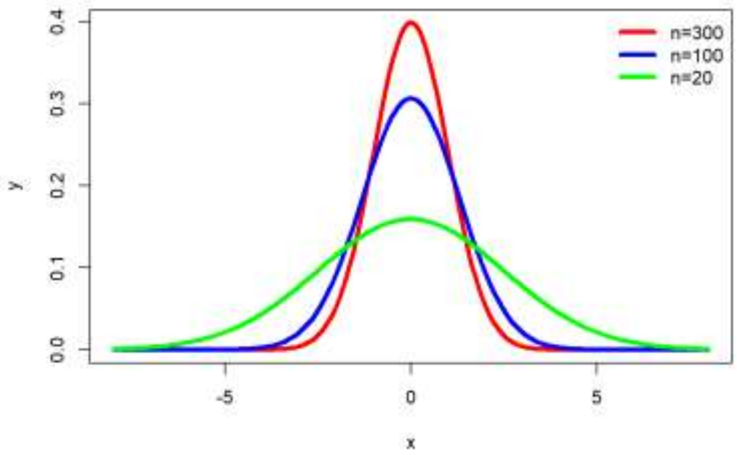
Standardizing a Normal Distribution and Z-scores

A distribution can be plotted using the raw scores found in the original data. That plot will have a mean and standard deviation calculated from the original data. To utilize the normal curve to determine probability functions and for inferential statistics we will want to convert that data so that it is standardized. We standardize so that the distribution is consistent across all distributions. That standardization produces a set of scores that have a mean of zero and a standard deviation of one. A standardized or Z-score of 1.5 means, therefore, that the score is one and a half standard deviations about the mean. A Z-score of -2.0 means that the score is two standard deviations below the mean.

As the formula above indicated, standardizing is a simple process. To move the mean from its original value to a mean of zero, all you have to do is subtract the mean from each score. To standardize the standard deviation to one all that is necessary is to divide each score the standard deviation.

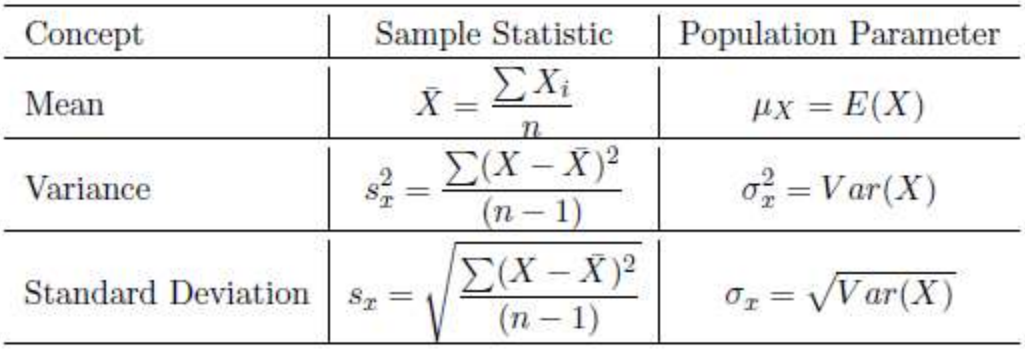
The Central Limit Theorem

An important property of samples is associated with the **Central Limit Theorem** (CLT). Imagine for a moment that we have a very large (or even infinite) population, from which we can draw as many samples as we'd like. According to the CLT, as the n-size (number of observations) within a sample drawn from that population increases, the more the distribution of the means taken from samples of that size will resemble a normal distribution. This is illustrated in the Figure below. Also note that the population does not need to have a normal distribution for the CLT to apply. Finally, a distribution of means from a normal population will be approximately normal at any sample size.



Populations, Samples and Symbols

It is important to note that, by convention, the symbols used for representing population parameters and sample statistics have different notation. These differences are shown in Table below. In short, population parameters are typically identified by using Greek letters and sample statistics are noted by English letters. Unless otherwise noted, the notation used in the remainder of this chapter will be in terms of samples rather than populations.



Concept	Sample Statistic	Population Parameter
Mean	$\bar{X} = \frac{\sum X_i}{n}$	$\mu_X = E(X)$
Variance	$s_x^2 = \frac{\sum(X - \bar{X})^2}{(n-1)}$	$\sigma_x^2 = Var(X)$
Standard Deviation	$s_x = \sqrt{\frac{\sum(X - \bar{X})^2}{(n-1)}}$	$\sigma_x = \sqrt{Var(X)}$

Inferences to the Population from the Sample

Another key implication of the Central Limit Theorem that is illustrated in the previous figure is that the mean of repeated sample means is the same, regardless of sample size, and that the mean of sample means is the population mean (assuming a large enough number of samples). Those conclusions lead to the important point that the sample mean is the best estimate of the population mean, i.e., the sample mean is an unbiased estimate of the population mean. The previous figure also illustrates as the sample size increases, the efficiency of the estimate increases. As the sample size increases, the mean of any particular sample is more likely to approximate the population mean.

When we begin our research we should have some population in mind - the set of items that we want to draw conclusions about. We might want to know about all adult Americans or about human beings (past, present, and future) or about a specific meteorological condition. There is only one way to know with certainty about that population and that is to examine all cases that fit the definition of our population. Most of the time, though, we cannot do that – in the case of adult Americans it would be very time-consuming, expensive, and logistically quite challenging, and in the other two cases it simply would be impossible. Our research, then, often forces us to rely on samples.

Because we rely on samples, inferential statistics are probability based. As the previous figure illustrates, our sample could perfectly reflect our population; it could be (and is likely to be) at least a reasonable approximation of the population; or the sample could deviate substantially from the population. Two critical points are being made here: the best estimates we have of our population parameters are our sample statistics, and we never know with certainty how good that estimate is. We make decisions (statistical and real world) based on probabilities.

Confidence Intervals

Because we are dealing with probabilities, if we are estimating a population parameter using a sample statistic, we will want to know how much confidence to place in that estimate. If we want to know a population mean, but only have a sample, the best estimate of that population mean is the sample mean. To know how much confidence to have in a sample mean, we put a confidence interval" around it. A confidence interval will report both a range for the estimate and the probability the population value falls in that range. We say, for example, that we are 95% confident that the true value is between A and B.

To find that confidence interval, we rely on the **standard error of the estimate**. The previous figure plots the distribution of sample statistics drawn from repeated samples. As the sample size increases, the estimates cluster closer to the true population value, i.e., the standard deviation is smaller. We could use the standard deviation from repeated samples to determine the confidence we can have in any particular sample, but in reality we are no more likely to draw repeated samples than we are to study the entire population. The standard error, though, provides an estimate of the standard deviation we would have if we had drawn a number of samples. The standard error is based on the sample size and the distribution of observations in our data:

$$SE = \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation, and n is the size (number of observations) of the sample.

The standard error can be interpreted just like a standard deviation. If we have a large sample, we can say that 68.26% of all of our samples (assuming we drew repeated samples) would fall within one standard error of our sample statistic or that 95.44% would fall within two standard errors.

If our sample size is not large, instead of using z-scores to estimate confidence intervals, we use **t-scores** to estimate the interval. *T*-scores are calculated just like z-score, but our interpretation of them is slightly different. The confidence interval formula is:

$$X \pm SE_x * t$$

To find the appropriate value for t, we need to decide what level of confidence we want (generally 95%) and our **degrees of freedom** (df), which is n−1. We can find a confidence interval with EXCEL using the regular math functions. Use the AVERAGE function to calculate your mean. Then calculate your standard error by first using the STDEV.S function to calculate s then divide by SQRT(n). Then multiple the SE by your t-values. For a 95% confidence interval, this is 1.96.

The Logic of Hypothesis Testing

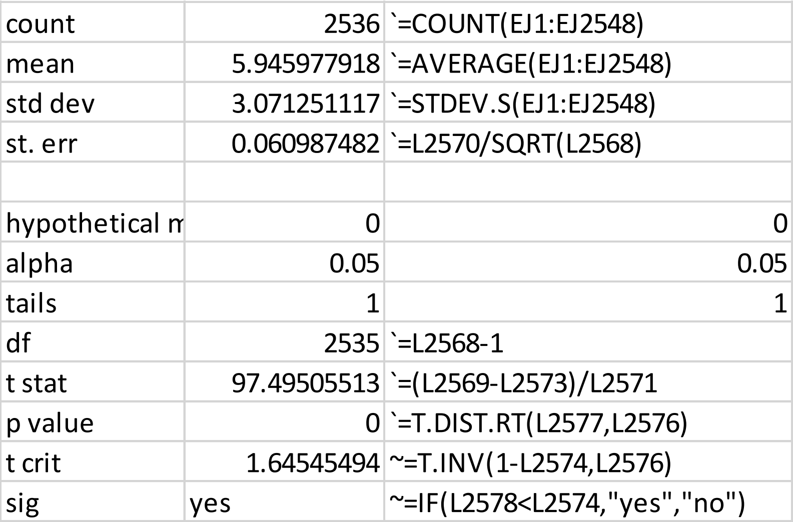
We can use the same set of tools to test hypotheses. In this section, we introduce the logic of hypothesis testing. In the next chapter, we address it in more detail. Remember that a **hypothesis** is a statement about the way the world is and that it may be true or false. Hypotheses are generally deduced from our theory and if our expectations are confirmed, we gain confidence in our theory. Hypothesis testing is where our ideas meet the real world.

Due to the nature of inferential statistics, we cannot directly test hypotheses, but instead we can test a **null hypothesis**. While a hypothesis is a statement of an expected relationship between two variables, the null hypothesis is a statement that says there is no relationship between the two variables. A null hypothesis might read:
As X increases, Y does not change. (We will discuss this topic more in the next chapter, but we want to understand the logic of the process here.)

Suppose a principal wants to cut down on absenteeism in her school and offers an incentive program for perfect attendance. Before the program, suppose the attendance rate was 85%. After having the new program in place for a while, she wants to know what the current rate is so she takes a sample of days and estimates the current attendance rate to be 88%. Her research hypothesis is: the attendance rate has gone up since the announcement of the new program (i.e., attendance is great than 85%). Her null hypothesis is that the attendance rate has not gone up since the announcement of the new program (i.e. attendance is less than or equal to 85%). At first it seems that her null hypothesis is wrong (88%>85%), but since we are using a sample, it is possible that the true population value is less than 85%. Based on her sample, how likely is it that the true population value is less than 85%? If the likelihood is small (and remember there will always be some chance), then we say our null hypothesis is wrong, i.e., we **reject our null hypothesis**, but if the likelihood is reasonable we accept our null hypothesis. The standard we normally use to make that determination is .05 – we want less than a .05 probability that we could have found our sample value (here 88%), if our null hypothesized value (85%) is true for the population. We use the t-statistic to find that probability. The formula is:

$$t = x - \frac{\mu}{se}$$

To test the hypothesis that our mean for risk perceptions of climate change (glbcc_risk) is different from zero in EXCEL you will need a workbook that looks like the following:



count	2536	`=COUNT(EJ1:EJ2548)
mean	5.945977918	`=AVERAGE(EJ1:EJ2548)
std dev	3.071251117	`=STDEV.S(EJ1:EJ2548)
st. err	0.060987482	`=L2570/SQRT(L2568)
hypothetical m	0	0
alpha	0.05	0.05
tails	1	1
df	2535	`=L2568-1
t stat	97.49505513	`=(L2569-L2573)/L2571
p value	0	`=T.DIST.RT(L2577,L2576)
t crit	1.64545494	~=T.INV(1-L2574,L2576)
sig	yes	~=IF(L2578<L2574,"yes","no")

The first column labels each row. The second is the output from the equations/functions that are typed in the third column. As you can see, our p-value is 0 (not technically, but rounded because it is so small) and therefore less that 0.05 and therefore significant. Meaning the mean of glbcc_risk which is 5.9 is different from 0.

Some Miscellaneous Notes about Hypothesis Testing

Before suspending our discussion of hypothesis testing, there are a few loose ends to tie up. First, you might be asking yourself where the .05 standard of hypothesis testing comes from. Is there some magic to that number? The answer is no"; .05 is simply the standard, but some researchers report .10 or .01. The p value of .05, though, is generally considered to provide a reasonable balance between making it nearly impossible to reject a null hypothesis and too easily cluttering our knowledge box with things that we think are related but actually are not. Even using the .05 standard means that 5% of the time when we reject the null hypothesis, we are wrong - there is no relationship. (Besides giving you pause wondering what we are wrong about, it should also help you see why science deems replication to be so important.)

Second, as we just implied, anytime we make a decision to either accept or reject our null hypothesis, we could be wrong. The probabilities tell us that if p=0.05, 5% of the time when we reject the null hypothesis, we are wrong because it is actually true. We call that type of mistake a **Type I Error**. However, when we accept the null hypothesis, we could also be wrong – there may be a relationship within the population. We call that a **Type II Error**. As should be evident, there is a trade-off between the two. If we decide to use a p value of .01 instead of .05, we make fewer Type I errors – just one out of 100, instead of 5 out of 100. Yet that also means that we increase by .04 the likelihood that we are accepting a null hypothesis that is false – a Type II Error. To rephrase the previous paragraph: .05 is normally considered to be a reasonable balance between the probability of committing Type I Errors as opposed to Type II Errors. Of course, if the consequence of one type of error or the other is greater, then you can adjust the p value.

Third, when testing hypotheses, we can use either a **one-tailed test** or a **two-tailed test**. The question is whether the entire .05 goes in one tail or is split evenly between the two tails (making, effectively, the p value equal to .025). Generally speaking, if we have a directional hypothesis (e.g., as X increases so does Y), we will use a one-tail test. If we are expecting a positive relationship, but find a strong negative relationship, we generally conclude that we have a sampling quirk and that the relationship is null, rather than the opposite of what we expected. If, for some reason, you have a hypothesis that does not specify the direction, you would be interested in values in either tail and use a two-tailed test.

Differences Between Groups

In addition to covariance and correlation (discussed in the next chapter), we can also examine differences in some variable of interest between two or more groups. For example, we may want to compare the mean of the perceived climate change risk variable for males and females. First, we can examine these variables visually.

As coded in our dataset, gender (gender) is a numeric variable with a 1 for male and 0 for female. To do this, we first need to remove non-valid responses from our data in EXCEL.

   a. Sort data from smallest to largest: Click on the Home tab on the toolbar at the top of your screen, select "Sort & Filter" on the far right side, then choose "Sort Smallest to Largest" from the drop-down menu that appears

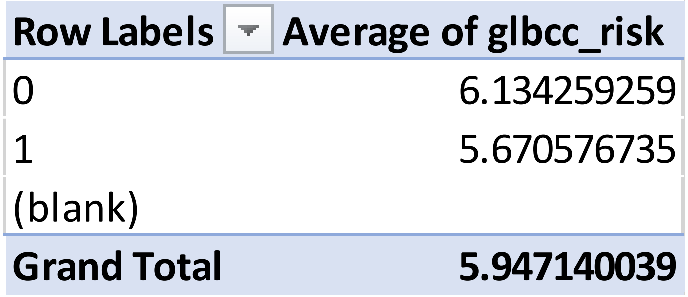
   b. Copy the segment with valid responses and paste into a new column or spreadsheet

      i. *Never delete the data! Always copy and paste when using this type of analysis so you do not lose valuable information you may need later. This is especially important if you are only using a few variables of a larger dataset.*

   c. Repeat this process for each column of data you are planning to use

2. Highlight the copied data to create a pivot table
3. Click on "Insert" tab on the toolbar at the top of the screen
4. Select "Pivot Table" on the far left side
5. Create pivot table on new worksheet
6. In the sidebar on the right side of your screen, click and drag the name of the independent variable to the quadrant labelled "rows" (just as you would for a means comparison table made on paper).
7. Click and drag the name of the dependent variable to the quadrant labelled "Values"
   a. Make sure the quadrant says "Average" of variable (the program will sometimes default to "Sum" of variable)
   b. To switch between these two, click on the small arrow next to the variable name and select "Value Field Settings…" at the bottom of the pop up menu
   c. "Average" is the third option in the list at the center of the dialog box
8. Click and drag the ID variable into the Values quadrant
   a. Make sure this variable is set to "Count" to determine how many observations fall into each category
   b. This should automatically change your "Columns" quadrant to "Σ Values"
9. If there is no control variable, this will result in all the information needed for your mean comparison table
   a. If you do have a control variable, add it to the Filters quadrant. This should add a small blue bar in the top left corner of the spread sheet.
   b. Click on the funnel icon of this bar to limit pivot table information to a single value of the control variable by de-selecting unneeded values. Repeat this process for each value of the control variable to fill in all columns of your mean comparison table.

For gender and climate change risk perceptions, you should end up with a table that looks like this:

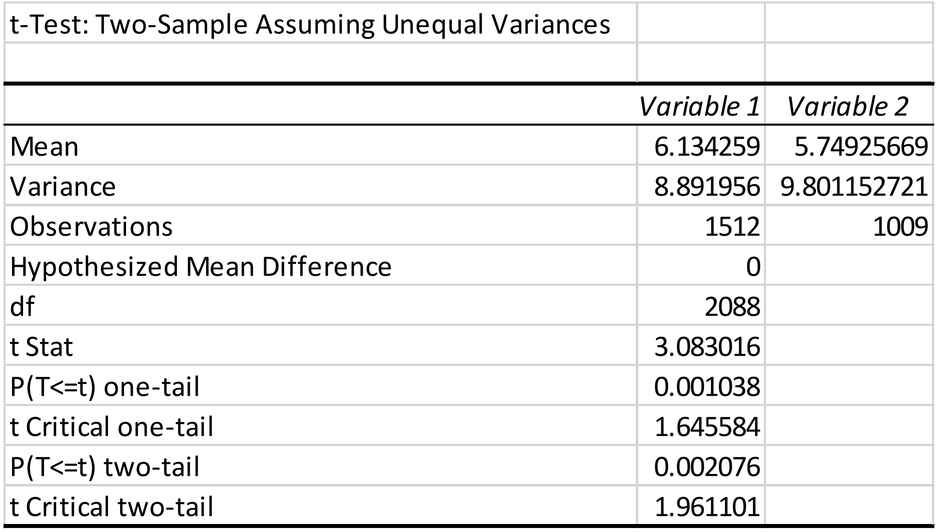


Row Labels	Average of glbcc_risk
0	6.134259259
1	5.670576735
(blank)	
Grand Total	**5.947140039**

The sheet created for this can also be used to conduct the corresponding t-test to tell if these means, 6.1 for Females and 5.6 for Males, are statistically different from each other. To do this, you simply need to have one column that is the risk perception variable for Females (0) and one for Males (1). Then:

1. Open Data Analysis ToolPak
2. Select t-Test: Two Sample Assuming Unequal Variances
   a. Unlikely to ever use equal variance option
   b. Paired t-test is for Before-After studies
3. Input appropriate arrays of cells into Variable 1 Range and Variable 2 Range from the columns you created in step 3
   a. Be sure to include the column titles in the areas you highlight
4. Hypothesized Mean Difference refers to the difference proposed by the null hypothesis, so input a zero here
5. Make sure the box next to "Labels" is checked
   a. If it is not checked, Excel will likely return an error message that your selected area includes non-numeric data
6. Alpha refers to level of significance, which automatically populates with 0.05
7. Select an output range (either new worksheet or highlighted range on current worksheet)
8. Click OK
   a. Test will return a table of information about the variables, with the values relevant to the t-test toward the bottom
      i. "t Stat" is the t-statistic just like we calculated by hand in class
      ii. "t Critical one-tail" and "t Critical two-tail" are the critical values against which your statistic is being compared, for one- and two-tailed tests
9. P(T<=t) refers to the probability value for one- or two-tailed tests
   a. If this value is less than the designated level of significance (likely 0.05), your relationship is significant and you can reject the null
   b. If this value is greater than the designated level of significance, your relationship is not significant and you cannot reject the null

This should create output in a new sheet that looks like this:



t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	6.134259	5.74925669
Variance	8.891956	9.801152721
Observations	1512	1009
Hypothesized Mean Difference	0	
df	2088	
t Stat	3.083016	
P(T<=t) one-tail	0.001038	
t Critical one-tail	1.645584	
P(T<=t) two-tail	0.002076	
t Critical two-tail	1.961101	

From this output, we can see that the p-value for both the one and two tail tests is less than 0.05 which suggests there is a statistically significant difference in means for climate change risk perceptions by gender. Specifically, males have lower risk perceptions than females. The null hypothesis is always that there is no relationship or difference. We test this hypothesis and if the p-value is less than our cut-off (0.05) then we reject the null! This is a double negative so it is a positive in result. We could thus phrase our finding as such: Therefore, we *reject the null hypothesis* and concluded that there are differences (on average) in the ways that males and females perceive climate change risk.

Summary

In this chapter we gained an understanding of inferential statistics, how to use them to place confidence intervals around an estimate, and an overview of how to use them to test hypotheses. In the next chapter we turn, more formally, to testing hypotheses using crosstabs and by comparing means of different groups. We then continue to explore hypothesis testing and model building using regression analysis.

Study Questions

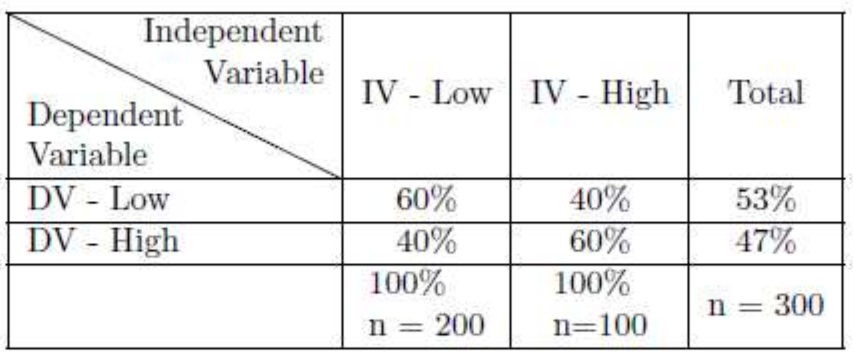
1. What is a hypothesis? What is a null hypothesis? Which one of these do we test?
2. Which type of error do we specify with our p-value () cut-offs?
3. Define probability sample; give at least two different examples of probability samples.
4. Why is rigorous and thoughtful sampling important? Give at least two reasons or examples of why analysts must carefully consider the type of sample for their research questions.

CHAPTER SEVEN: Association of Variables

The last chapter focused on the characterization of distributions of a single variable. We now turn to the associations between two or more variables. This chapter explores ways to measure and visualize associations between variables. We start with how to analyze the relations between nominal and ordinal level variables, using cross-tabulation in EXCEL. Then, for interval level variables, we examine the use of the measures of covariance and correlation between pairs of variables. Next, we examine hypothesis testing between two groups, where the focus in on how the groups differ, on average, with respect to an interval level variable. Finally, we discuss scatterplots as a way to visually explore differences between pairs of variables.

Cross-Tabulation

To determine if there is an association between two variables measured at the nominal or ordinal levels, we use cross-tabulation and a set of supporting statistics. A cross-tabulation (or just crosstab) is a table that looks at the distribution of two variables simultaneously. Table below provides a sample layout of a 2 X 2 table.



Independent Variable Dependent Variable	IV - Low	IV - High	Total
DV - Low	60%	40%	53%
DV - High	40%	60%	47%
	100% n = 200	100% n=100	n = 300

As Table above illustrates, a crosstab is set up so that the independent variable is on the top, forming columns, and the dependent variable is on the side, forming rows. Toward the upper left hand corner of the table are the low, or negative, variable categories. Generally, a table will be displayed in percentage format. The marginals for a table are the column totals and the row totals and are the same as a frequency distribution would be for that variable. Each cross-classification reports how many observations have that shared characteristic. The cross-classification groups are referred to as **cells**, so Table above is a four-celled table.

A table like Table above provides a basis to begin to answer the question of whether our independent and dependent variables are related. Remember that our null hypothesis says there is no relationship between our IV and our DV. Looking at Table above , we can say of those low on the IV, 60% of them will also be low on the DV; and that those high on the IV will be low on the DV 40% of the time. Our null hypothesis says there should be no difference, but in this case, there is a 20% difference so it appears that our null hypothesis is incorrect. What we learned in our inferential statistics chapter, though, tells us that it is still possible that the null hypothesis is true. The question is how likely is it that we could have a 20% difference in our sample even if the null hypothesis is true?[1]

We use the **chi square statistic** to test our null hypothesis when using crosstabs. To find chi square (χ2χ2), we begin by assuming the null hypothesis to be true and find the expected frequencies for each cell in our table. We do so using a posterior methodology based on the marginals for our dependent variable. We see that 53% of our total sample is low on the dependent variable. If our null hypothesis is correct, then where one is located on the independent variable should not matter: 53% of those who are low on the IV should be low on the DV and 53% of those who are high on the IV should be low on the DV. Tables with the Null Hypothesis as Percentages and as Counts illustrate this pattern. To find the expected frequency for each cell, we simply multiply the expected cell percentage times the number of people in each category of the IV: the expected frequency for the low-low cell is .53∗200=106=; for the low-high cell, it is .47∗200=94=; for the low-high cell it is .53∗100=53; and for the high-high cell, the expected frequency is .47∗100=47.

The formula for the chi square takes the expected frequency for each of the cells and subtracts the observed frequency from it, squares those differences, divides by the expected frequency, and sums those values:

$$\mathrm{X}^2 = \Sigma \frac{(O - E)^2}{E}$$

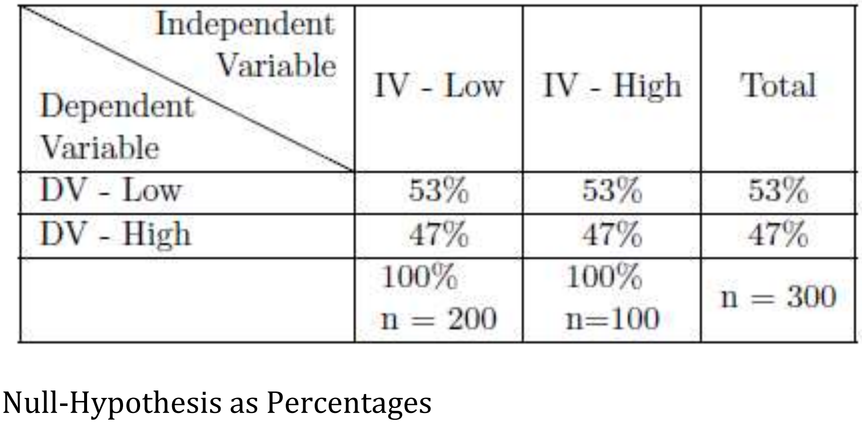
where:

χ2 = The Test Statistic

∑ = The Summation Operator

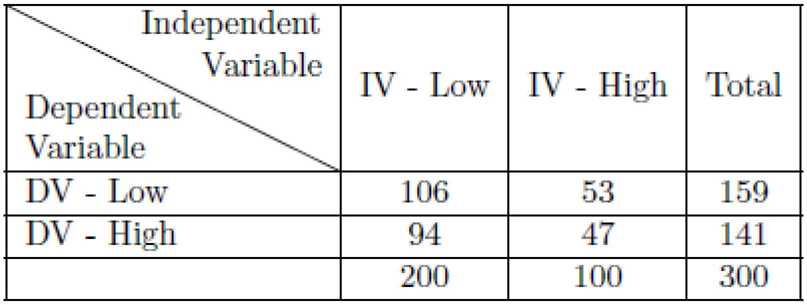
O = Observed Frequencies

E = Expected Frequencies



Independent Variable Dependent Variable	IV - Low	IV - High	Total
DV - Low	53%	53%	53%
DV - High	47%	47%	47%
	100% n = 200	100% n=100	n = 300

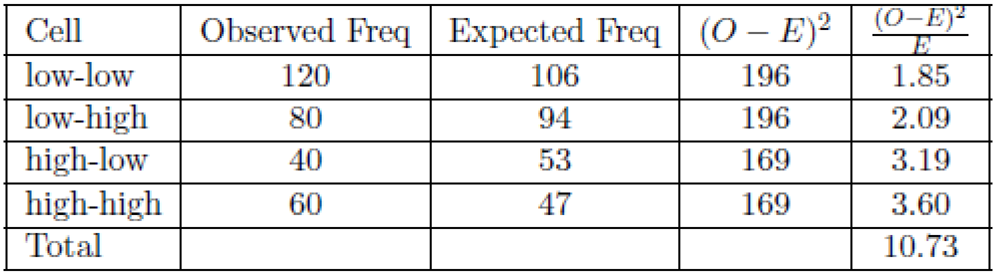
Null-Hypothesis as Percentages



Independent Variable Dependent Variable	IV - Low	IV - High	Total
DV - Low	106	53	159
DV - High	94	47	141
	200	100	300

Null Hypothesis as Counts

The table below provides those calculations. It shows a final chi square of 10.73. With that chi square, we can go to a chi square table to determine whether to accept or reject the null hypothesis. Before going to that chi square table, we need to figure out two things. First, we need to determine the level of significance we want, presumably .05. Second, we need to determine our degrees of freedom. We will provide more on that concept as we go on, but for now, know that it is the number of rows minus one times the number of columns minus one. In this case we have (2−1)(2−1)=1 degree of freedom.



Cell	Observed Freq	Expected Freq	$(O-E)^2$	$\frac{(O-E)^2}{E}$
low-low	120	106	196	1.85
low-high	80	94	196	2.09
high-low	40	53	169	3.19
high-high	60	47	169	3.60
Total				10.73

The Table at the end of this chapter is a chi square table that shows the critical values for various levels of significance and degrees of freedom. The critical value for one degree of freedom with a .05 level of significance is 3.84. Since our chi square is larger than that we can reject our null hypothesis - there is less than a .05 probability that we could have found the results in our sample if there is no relationship in the population. In fact, if we follow the row for one degree of freedom across, we see we can reject our null hypothesis even at the .005 level of significance and, almost but not quite, at the .001 level of significance.

Having rejected the null hypothesis, we believe there is a relationship between the two variables, but we still want to know how strong that relationship is. Measures of association are used to determine the strength of a relationship. One type of measure of association relies on a co-variation model as elaborated upon in previous sections. Co-variation models are directional models and require ordinal or interval level measures; otherwise, the variables have no direction. Here we consider alternative models.

If one or both of our variables is nominal, we cannot specify directional change. Still, we might see a recognizable pattern of change in one variable as the other variable varies. Women might be more concerned about climate change than are men, for example. For that type of case, we may use a reduction in error or a **proportional reduction in error (PRE)**

model. We consider how well we predict using a naive model (assuming no relationship) and compare it to how much better we predict when we use our independent variable to make that prediction. These measures of association only range from 0−1.0, since the sign otherwise indicates direction. Generally, we use this type of measure when at least one our variables is nominal, but we will also use a PRE model measure, R^2, in regression analysis. **Lambda** is a commonly used PRE-based measure of association for nominal level data, but it can underestimate the relationship in some circumstances.

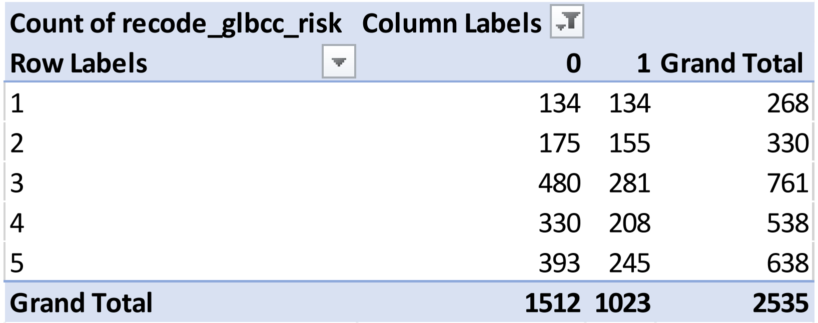
Another set of measures of association suitable for nominal level data is based on chi square. **Cramer's V** is a simple chi square based indicator, but like chi square itself, its value is affected by the sample size and the dimensions of the table. **Phi** corrects for sample size, but is appropriate only for a 2 X 2 table. The **contingency coefficient**, C, also corrects for sample size and can be applied to larger tables, but requires a square table, i.e., the same number of rows and columns.

If we have ordinal level data, we can use a co-variation model, but the specific model developed below in Section 6.3 looks at how observations are distributed around their means. Since we cannot find a mean for ordinal level data, we need an alternative. **Gamma** is commonly used with ordinal level data and provides a summary comparing how many observations fall around the diagonal in the table that supports a positive relationship (e.g. observations in the low-low cell and the high-high cells) as opposed to observations following the negative diagonal (e.g. the low-high cell and the high-low cells). Gamma ranges from −1.0to +1.0.

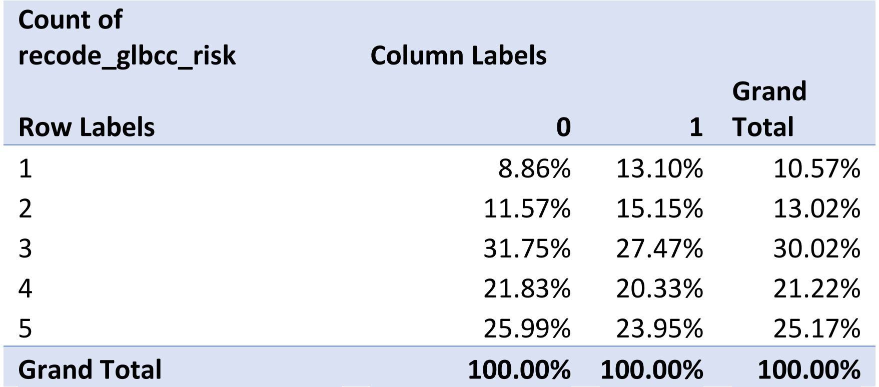
Crosstabulations and their associated statistics can be calculated using EXCEL. In this example we continue to use the Global Climate Change dataset (ds). The dataset includes measures of survey respondents: gender (female = 0, male = 1); perceived risk posed by climate change, or glbcc_risk (0 = Not Risk; 10 = extreme risk), and political ideology (1 = strong liberal, 7 = strong conservative). Here we look at whether there is a relationship between gender and the glbcc_risk variable. The glbcc_risk variable has eleven categories; to make the table more manageable, we recode it to five categories. This recode can be done using =IF() commands in EXCEL. Specifically, if we want 0-1 to now = 1, 2-3 to = 2, 4-6 to =3, 7-8 =4, and 9-10 to = 5 we need this function:

```
=IF(C2<2,1,IF(OR(C2=2,C2=3),2,IF(AND(C2>=4,C2<=6),3,IF(OR(C2=7,C2=8),4,IF(OR(C2=9,
C2=10),5,0)))))
```

We can use similar Pivot Tables to those taught elsewhere to get this (rows: recoded variable; columns: gender; values: count of recoded – changing this last option to % of column would also be useful and is the next table).



Count of recode_glbcc_risk	Column Labels		
Row Labels	0	1	Grand Total
1	134	134	268
2	175	155	330
3	480	281	761
4	330	208	538
5	393	245	638
Grand Total	**1512**	**1023**	**2535**

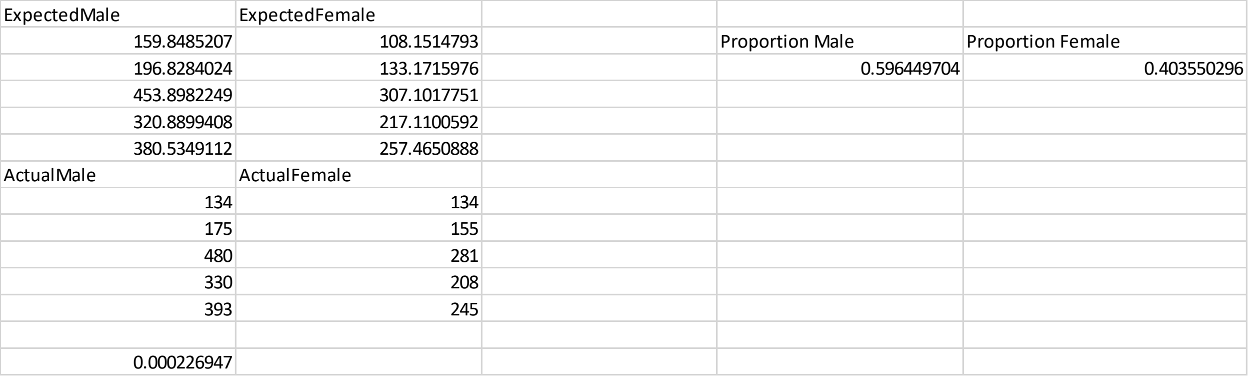


Count of recode_glbcc_risk	Column Labels		
Row Labels	0	1	Grand Total
1	8.86%	13.10%	10.57%
2	11.57%	15.15%	13.02%
3	31.75%	27.47%	30.02%
4	21.83%	20.33%	21.22%
5	25.99%	23.95%	25.17%
Grand Total	**100.00%**	**100.00%**	**100.00%**

To actually test for a relationship, follow the below steps:

1. Create a cross-tab using the pivot table method
2. Calculate expected frequencies for each cell
3. Translate into a table of observed and expected frequencies
    a. This table should be outside of pivot table layout in standalone cells
    b. Result will look similar to a cross-tab, with IV categories in the columns and DV categories in the rows
    c. Need two copies of each category of the DV in this table (one for actual and one for expected)
        i. It does not matter which is placed on top, either way will work (formula returns same result either way)
        ii. Make sure they are appropriately labelled so you know which is which
4. Insert the "CHISQ.TEST" formula
    a. Highlight appropriate arrays for actual and expected range
        i. Actual range is observed frequencies portion of the table
        ii. Expected range is expected frequencies portion of the table
5. Formula will return a single number—this is your p-value
    a. If less than designated significance level (most likely 0.05), chi square is statistically significant and you can reject the null
    b. If greater than designated significance level, chi square is not statistically significant and you cannot reject the null

The following set up can be used to test this hypothesis. The actual counts are from the count Pivot Table above. The Expected counts are calculated by multiplying the proportion by the sum of the actual so: 0.596 * (175+155) will give us the expected count for Male with the second level of risk and so on. In this case, that bottom cell is our p-value. The CHISQ.TEST function is used to return this. The first argument is the array of actual values and the second is array of expected values.



ExpectedMale	ExpectedFemale			
159.8485207	108.1514793		Proportion Male	Proportion Female
196.8284024	133.1715976		0.596449704	0.403550296
453.8982249	307.1017751			
320.8899408	217.1100592			
380.5349112	257.4650888			
ActualMale	ActualFemale			
134	134			
175	155			
480	281			
330	208			
393	245			
0.000226947				

In this case, we return a p-value less that 0.05 therefore we reject the null hypothesis. Thus, we have evidence that there is a relationship between gender and climate change risk perceptions.

Covariance

Covariance is a simple measure of the way two variables move together, or "co-vary". The covariance of two variables, X and Y, can be expressed in population notation as:

$$cov(X,Y) = E\left[(X-\mu_x)\left(Y-\mu_y\right)\right]$$

Therefore, the covariance between X and Y is simply the product of the variation of X around its expected value, and the variation of Y around its expected value. The sample covariance is expressed as:

$$cov(X,Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1}$$

Covariance can be positive, negative, or zero. If the covariance is positive *both variables move in the same direction*, meaning if X increases Y increases or if X decreases Y decreases. Negative covariance means that the *variables move in opposite directions*; if X increases Y decreases. Finally, zero covariance indicates that there is no covariance between X and Y.

Correlation

Correlation is closely related to covariance. In essence, correlation standardizes covariance so it can be compared across variables. Correlation is represented by a correlation coefficient, ρ, and is calculated by dividing the covariance of the two variables by the product of their standard deviations. For populations it is expressed as:

$$\rho = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

For samples it is expressed as:

$$r = \frac{\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1}}{s_x s_y}$$

Like covariance, correlations can be positive, negative, and zero. The possible values of the correlation coefficient r, range from -1, perfect negative relationship to 1, perfect positive relationship. If r=0, that indicates no correlation. Correlations can be calculated in EXCEL, using a few different methods.

Option 1

1. In an empty cell, insert the function CORREL or PEARSON. These can be found by searching "correlation" in the "Insert Function" dialog box or in the Statistical functions category under the "More Functions" menu.
    a. Remember both of these functions are performing the same calculation, so you will only need to use one of them.
2. This should cause a new dialog box to pop up on your screen.
3. In the box labeled "Array 1," highlight the label and data for your first variable. In the box labeled "Array 2," highlight the label and data for your second variable.
    a. It does not matter which variable goes into which box, as correlation is an association measure rather than a causal one. It will return the same answer either way.
    b. Make sure you do not include any additional information in this range for either variable, such as descriptive statistics of the variable, as this will cause the formula to return incorrect results.
4. Click OK or press Enter on the keyboard to complete the function.
5. The number in the cell is your correlation coefficient (r).

Or

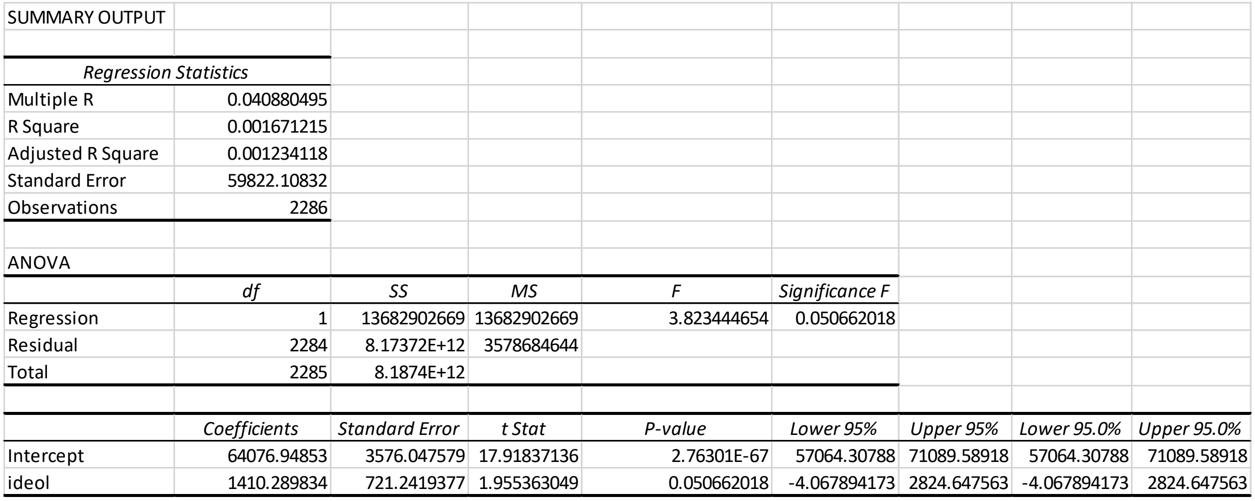
Option 2

1. On the Data tab of the toolbar at the top of the screen, select Data Analysis to open the ToolPak dialog box.
2. Scroll through the list, select Correlation, and click OK.
3. In the "Input Range" box, select all of the data you want to correlate. The range should include both variables and their labels in row 1.
    a. This function will also allow you to select more than two variables if you need to do so.
4. Make sure the box next to "Labels in First Row" is checked.
5. Select an output range for the correlation table (the default is to create a new sheet).
6. Click OK to complete the function and generate the correlation table.
7. The correlation coefficient (r) for the two variables is listed in the box at the intersection of the two variable names.

In order to get a p-value so that we can say something about the statistical significance of a correlation, we must follow other steps. This is important because this lets us compare the correlation to a zero to tell us if the correlation is "real" that is statistically different from zero or not. Follow these steps:

1. Open the Data Analysis ToolPak (located on the Data tab of the toolbar).
2. Scroll through the list of options, select Regression, and click OK.
3. In the "Input Y Range" box, highlight the first variable you are using, including its label in the first row. In the "Input X Range" box, highlight the second variable you are using, including its label in the first row.
    a. It does not matter which variable goes into which box, as correlation is an association measure rather than a causal one. It will return the same answer either way.
    b. Make sure you do not include any additional information in this range for either variable, such as descriptive statistics of the variable, as this will cause the formula to return incorrect results.
4. Check the box next to "Labels."
5. Choose your output location for the results (the default is to create a new sheet).
6. Click OK to complete the function, which should generate several tables.
7. The p-value of the correlation is listed in two places:
    a. In the second table, labelled "ANOVA" the p-value is listed in the rightmost column, "Significance F."
    b. In the third table (at the bottom of the output), the p-value is listed in the fifth column from the left, titled P-value. The number you are looking for will be the bottom row of the table, and should be the same as the number in the Significance F column of the ANOVA table.
8. Compare this p-value to the significance level (0.05) to determine statistical significance.

We can do this for income and ideology and get the following results.

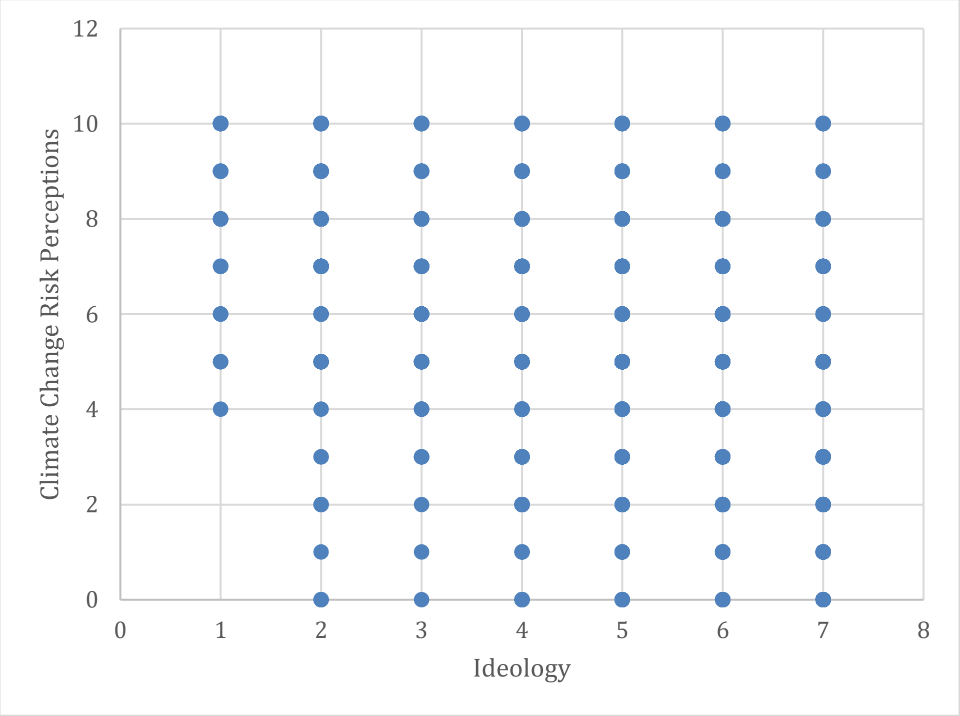


SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.040880495							
R Square	0.001671215							
Adjusted R Square	0.001234118							
Standard Error	59822.10832							
Observations	2286							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	13682902669	13682902669	3.823444654	0.050662018			
Residual	2284	8.17372E+12	3578684644					
Total	2285	8.1874E+12						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	64076.94853	3576.047579	17.91837136	2.76301E-67	57064.30788	71089.58918	57064.30788	71089.58918
ideol	1410.289834	721.2419377	1.955363049	0.050662018	-4.067894173	2824.647563	-4.067894173	2824.647563

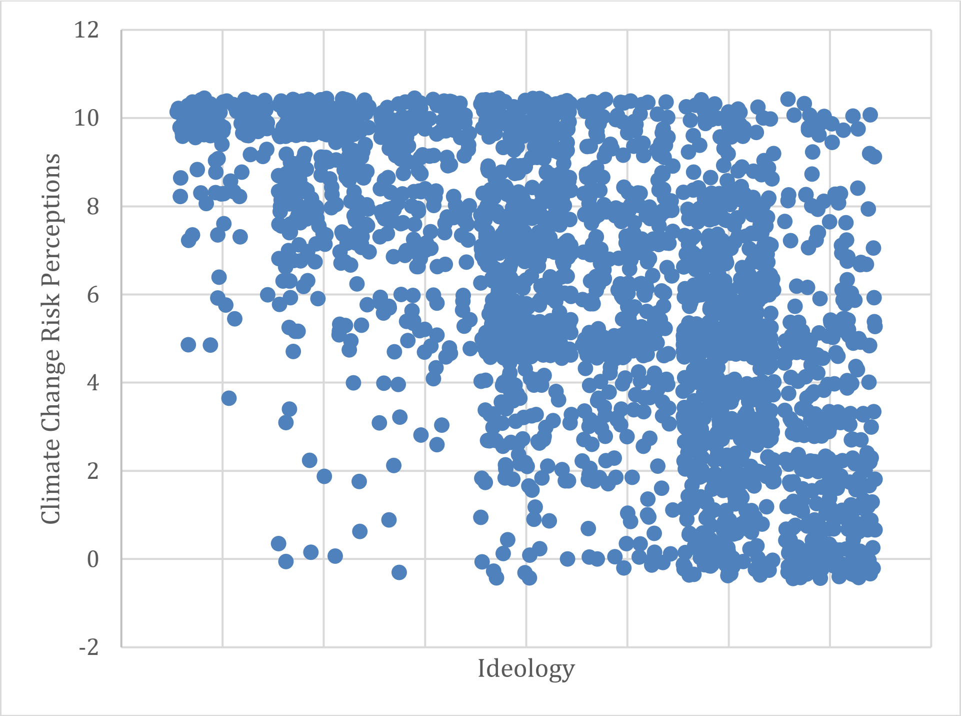
The multiple R row will tell us the correlation that should be equal to if we use one of the first methods (=CORREL or =PEARSON). Then the p-value to interpret is described in the steps above. We can see in this example the p-value is 0.0506. This p-value is NOT less than 0.05 so we FAIL to reject the null hypothesis. Thus, we find no statistically significant correlation between income and ideology.

Scatterplots

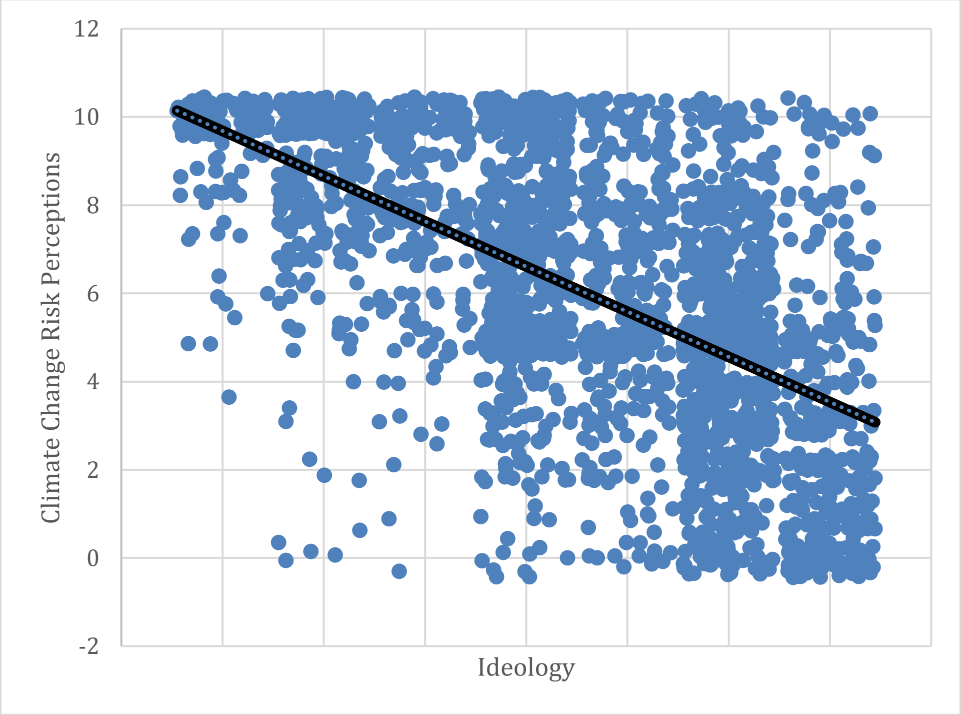
As noted earlier, it is often useful to try and see patterns between two variables. We examined the density plots of males and females with regard to climate change risk, then we tested these differences for statistical significance. However, we often want to know more than the mean difference between groups; we may also want to know if differences exist for variables with several possible values. For example, here we examine the relationship between ideology and perceived risk of climate change. This can be done using the Insert, Charts, and Scatter option which is the last of the small icons. Below is the relationship between ideology (7= Strongly Conservative) and climate change risk perceptions.



When we look at this, it isn't very useful because our data points are all plotted over each other. In EXCEL, we can jitter our points using the following function: A2 + ((RAND() - 0.5) * 0.9. This adds a small random amount to our data then subtracts 0.5 to sometimes make the small amount added negative and then scales it. In this case, we use 0.9 because the interval between our actual data is equal to 1. The resulting figure is below:

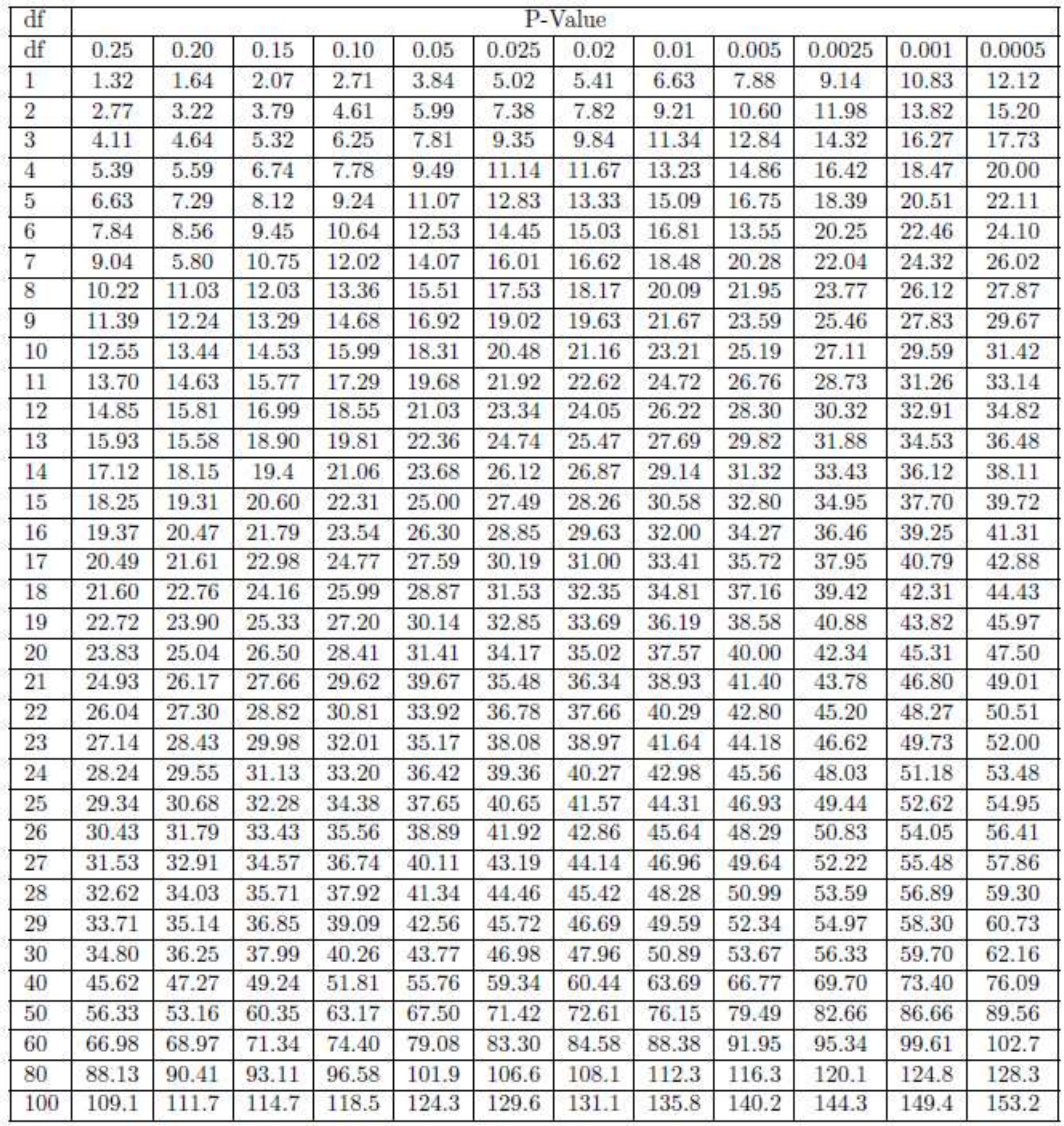


This figure is much better and suggests a potential negative relationship between ideology and climate change risk perceptions. Conservatives have lower risk perceptions of climate change. We can also plot a line over this that we will eventually calculate in the coming chapters and in fact have already learned with the Regression command from the ToolPak:



This is done through the add chart element Linear Trend and then formatting the line to be solid and black so that it can be seen over the blue points. Note that the regression lines both slope downward, with average perceived risk ranging from over 8 for the strong liberals (ideology=1) to less than 5 for strong conservatives (ideology=7). This illustrates how scatterplots can provide information about the nature of the relationship between two variables. We will take the next step – to bivariate regression analysis – in the next chapter.

End of Chapter Chi-Square Table referenced ear



df	P-Value											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.59	6.74	7.78	9.49	11.14	11.67	13.23	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.33	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.53	14.45	15.03	16.81	13.55	20.25	22.46	24.10
7	9.04	5.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.63	21.67	23.59	25.46	27.83	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.29	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.93	15.58	18.90	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.4	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	39.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	53.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

Study Questions

1. What is the first step in any association of variables analysis?
2. Chi-square statistics are used for assessing the existence of a relationship in cross-tabs. This method is therefore most useful for variables of what level of measurement?
3. What is the range of possible values for correlation? Explain what a negative, positive, and zero correlation mean in.
4. Correlation does NOT imply causation. Why? Why are correlations still very important in social science research?

CHAPTER EIGHT: The Logic of Ordinary Least Squares Estimation

This chapter begins the discussion of ordinary least squares (OLS) regression. OLS is the "workhorse" of empirical social science and is a critical tool in hypothesis testing and theory building. This chapter builds on the discussion in Chapter 6 by showing how OLS regression is used to estimate relationships between and among variables.

Theoretical Models

Models, as discussed earlier, are an essential component in theory building. They simplify theoretical concepts, provide a precise way to evaluate relationships between variables, and serve as a vehicle for hypothesis testing. As discussed in Chapter 1, one of the central features of a theoretical model is the presumption of causality, and causality is based on three factors: time ordering (observational or theoretical), co-variation, and non-spuriousness. Of these three assumptions, co-variation is the one analyzed using OLS. The oft repeated adage, 'correlation is not causation' is key. Causation is driven by theory, but co-variation is the critical part of empirical hypothesis testing.

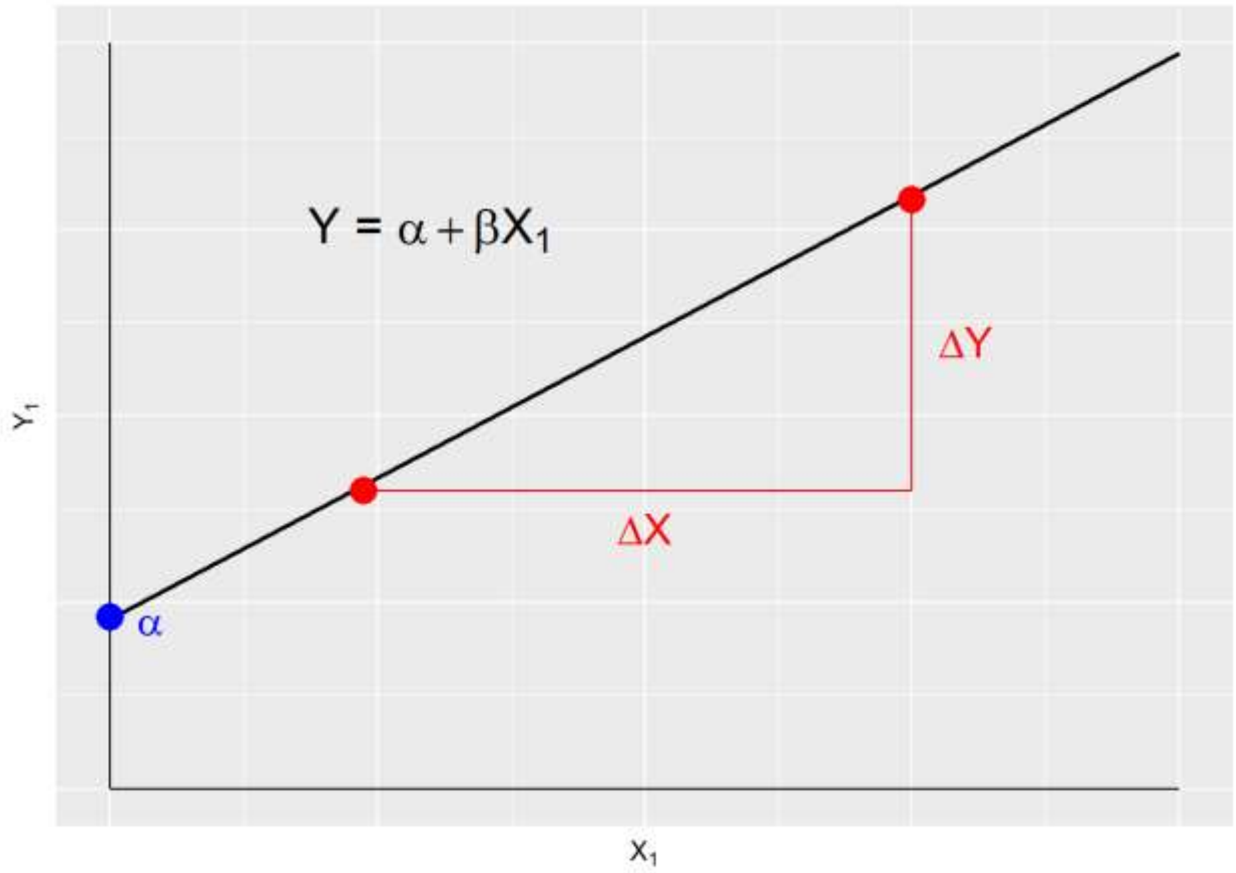
When describing relationships, it is important to distinguish between those that are *deterministic* versus *stochastic*. Deterministic relationships are "fully determined" such that, knowing the values of the independent variable, you can perfectly explain (or predict) the value of the dependent variable. Philosophers of Old (like Kant) imagined the universe to be like a massive and complex clock which, once wound up and set ticking, would permit perfect prediction of the future if you had all the information on the starting conditions. There is no "error" in the prediction. Stochastic relationships, on the other hand, include an irreducible random component, such that the independent variables permit only a partial prediction of the dependent variable. But that stochastic (or random) component of the variation in the dependent variable has a probability distribution that can be analyzed statistically.

Deterministic Linear Model

The deterministic linear model serves as the basis for evaluating theoretical models. It is expressed as:

$Y_i = \alpha + \mathrm{B}X_i$

A deterministic model is **systematic** and contains no error therefore Y is perfectly predicted by X. This is illustrated in the figure below α and β are the model parameters, and are constant terms. β is the slope, or the change in Y over the change in X. α is the intercept, or the value of Y when X is zero.



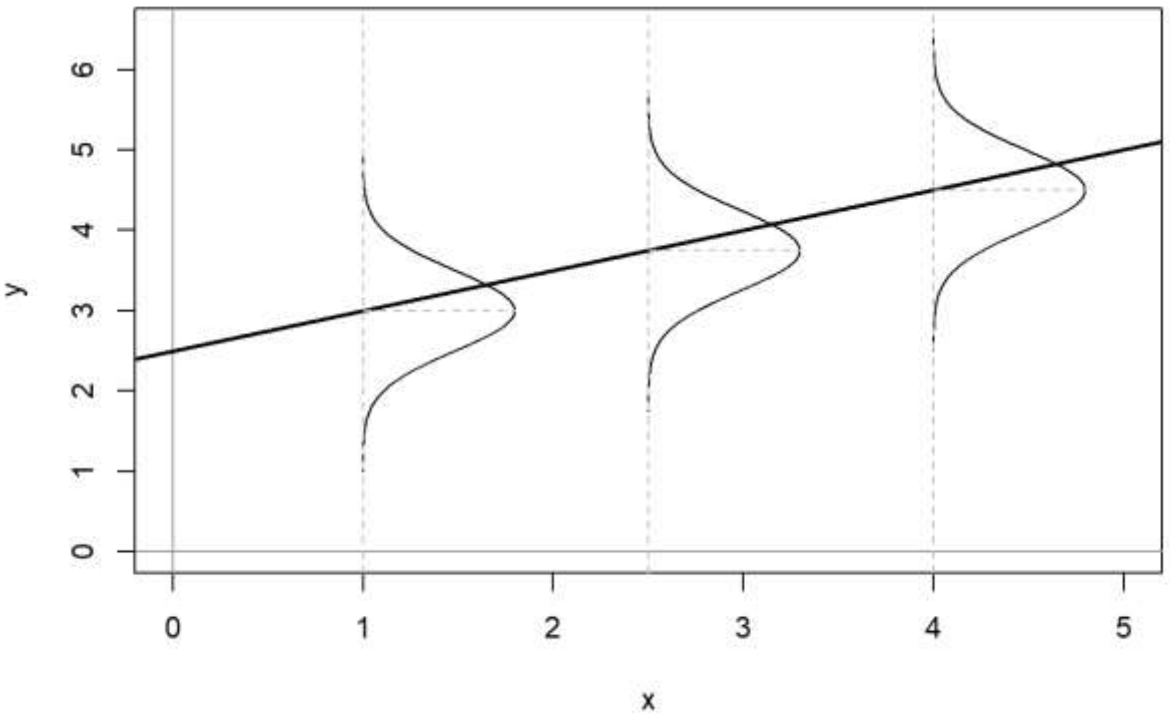
Given that in social science we rarely work with deterministic models, nearly all models contain a stochastic, or random, component.

Stochastic Linear Model

The stochastic, or statistical, linear model contains a systematic component, Y=α+β and a stochastic component called the **error term**. The error term is the difference between the expected value of Yi and the observed value of Yi; Yi−μ. This model is expressed as:

$Y_i = \alpha + \mathrm{B}X_i + \epsilon_i$

where ϵi is the error term. In the deterministic model, each value of Y fits along the regression line, however in a stochastic model the expected value of Y is conditioned by the values of X. This is illustrated in the Figure below.



Assumptions about the Error Term

There are three key assumptions about the error term; a) errors have identical distributions, b) errors are independent, and c) errors are normally distributed.[1]

Error Assumptions

- Errors have identical distributions
  - $E(\epsilon_i^2) = \sigma_i^2$
- Errors are independent of X and other ϵ_i
  - $E(\epsilon_i) = E(\epsilon|x_i) = 0$

  And

  - $(\epsilon_i) \neq E(\epsilon_j), for\ i \neq j$
- Errors are normally distributed
  - $\epsilon_i \sim N(0, \sigma_i^2)$

Taken together these assumptions mean that the error term has a normal, independent, and identical distribution (normal i.i.d.). However, we don't know if, in any particular case, these assumptions are met. Therefore, we must estimate a linear model.

Estimating Linear Models

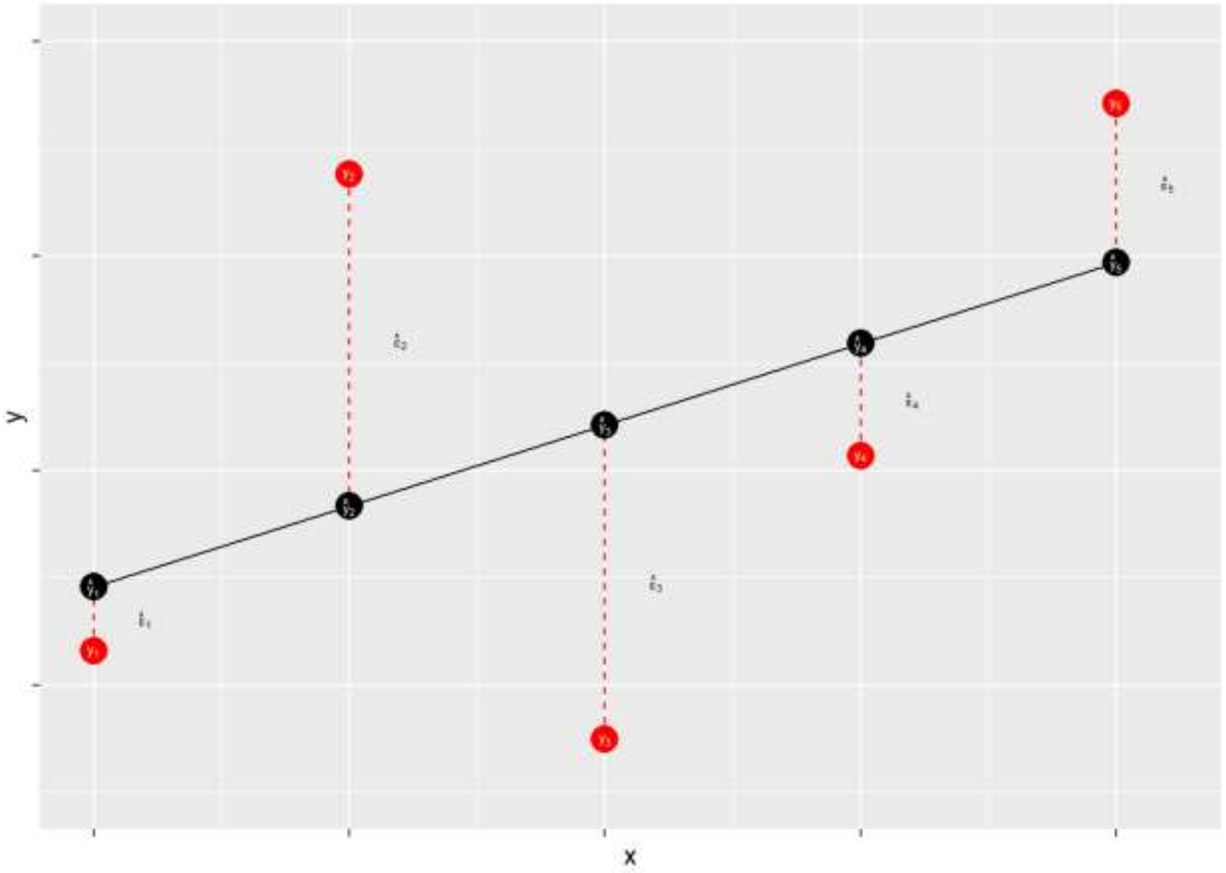
With stochastic models we don't know if the error assumptions are met, nor do we know the values of α and β; therefore we must estimate them, as denoted by a hat (e.g., $\widehat{\alpha}$ is the estimate for α). The stochastic model as shown in the equation below is estimated as:

$$Y_i \ = \ \hat{\alpha} + \ \widehat{\mathrm{B}}X_i + \ \epsilon_i$$

where ϵ_i is the **residual term**, or the estimated error term. Since no line can perfectly pass through all the data points, we introduce a residual, ϵ, into the regression equation. Note that the predicted value of Y is denoted $\hat{Y}$.

Residuals

Residuals measure prediction errors of how far observation Yi is from predicted $\hat{Y}$. This is shown in the figure below.



The residual term contains the accumulation (sum) of errors that can result from measurement issues, modeling problems, and irreducible randomness. Ideally, the residual term contains lots of small and independent influences that result in an overall random quality of the distribution of the errors. When that distribution is not random – that is, when the distribution of error has some systematic quality – the estimates of $\hat{\alpha}\ and\ \widehat{\mathrm{B}}$ may be biased. Thus, when we evaluate our models we will focus on the shape of the distribution of our errors.

What's in ϵ?

Measurement Error

- Imperfect operationalizations
- Imperfect measure application

Modeling Error

- Modeling error/mis-specification
- Missing model explanation
- Incorrect assumptions about associations
- Incorrect assumptions about distributions

Stochastic "noise"

- Unpredictable variability in the dependent variable

The goal of regression analysis is to minimize the error associated with the model estimates. As noted, the residual term is the estimated error, or overall miss" (e.g., Yi− Y). Specifically, the goal is to minimize the sum of the squared errors, $\sum\epsilon^2$. Therefore, we need to find the values of $\hat{\alpha}\ and\ \widehat{\mathrm{B}}$ that minimize $\sum\epsilon^2$.

Note that for a fixed set of data each possible choice of values for $\hat{\alpha}\ and\ \widehat{\mathrm{B}}$ corresponds to a specific residual sum of squares, $\sum\epsilon^2$. This can be expressed by the following functional form:

$$S(\hat{\alpha}\,,\widehat{\mathrm{B}}) = \sum_{i=1}^{n}\epsilon_i^2 = \sum\left(Y_i - \hat{Y}_\iota\right)^2 = \sum(Y_i - \ \hat{\alpha} - \ \widehat{\mathrm{B}}X_i)^2$$

Minimizing this function requires specifying estimators for $\hat{\alpha}\ and\ \widehat{\mathrm{B}}$ such that $S(\hat{\alpha}\,,\widehat{\mathrm{B}}) = \Sigma\epsilon^2$ is at the lowest possible value. Finding this minimum value requires the use of calculus, which will be discussed in the next chapter. Before that we walk through a quick example of simple regression.

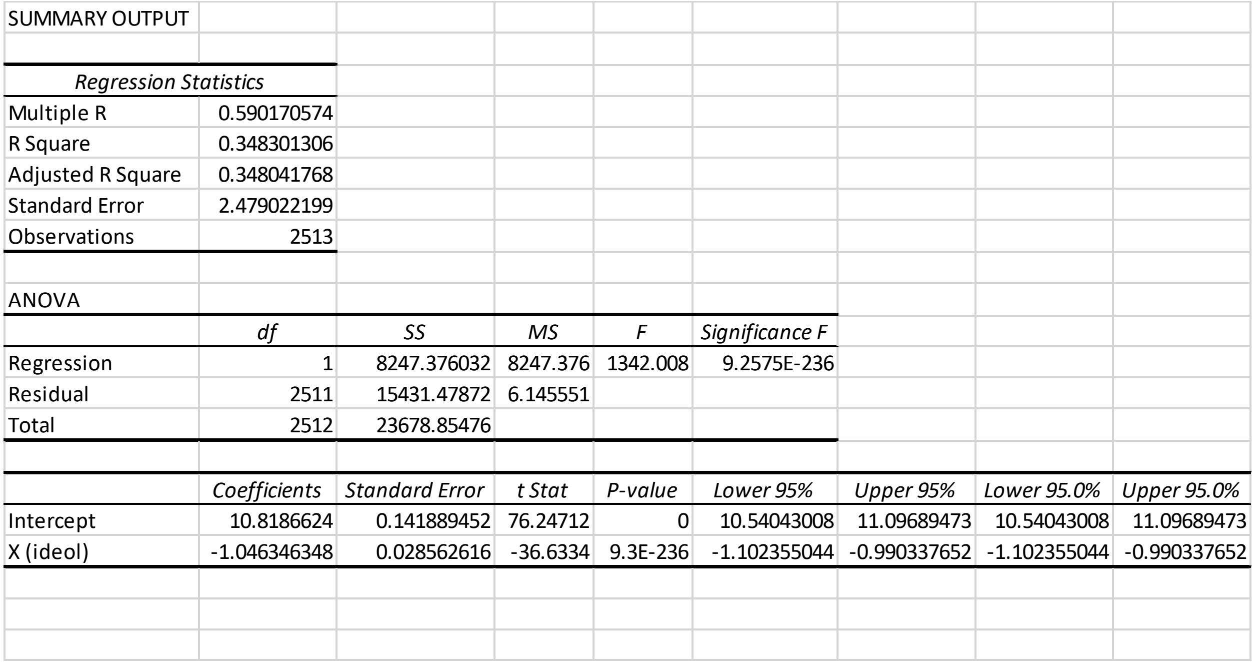
An Example of Simple Regression

The following example uses a measure of peoples' political ideology to predict their perceptions of the risks posed by global climate change. OLS regression can be done using the Data Analysis ToolPak in Excel following these steps:

1. Select the variables you will be using for your analysis. It is best to copy them to a new spreadsheet or file before you begin working with them.
    a. If you are performing a multivariate analysis, make sure all independent variables are in columns next to each other, with no gaps or other data in between.
2. Remove any non-valid responses (i.e., "don't know", "not applicable", etc.) from your data before performing any statistical analysis. The easiest way to do this is by sorting each variable and deleting the rows containing non-valid measurements.
3. On the toolbar at the top of the screen, click the Data tab.
4. To the far right of the screen should be a link to Data Analysis. Clicking this will open the Toolpak dialog box.
5. Select Regression from the list of Analysis Tools and click OK to open a dialog box for the regression input.
6. In the Input Y Range box, highlight your dependent variable. Select only the rows that contain data, as the function will not run if the entire column of the spreadsheet is selected. (E.g., If you have 2,841 observations, your selection should only contain 2,842 rows—one for each observation and the title row.)
    a. It is extremely important that you put the variables into the right boxes at this stage, or the function will return the wrong results. This box should only ever have one variable in it.
    b. Be sure to include the column title in your selection.
7. In the Input X Range box, highlight the columns containing your independent variable(s) and any control variables you are including in your analysis. Select only the rows that contain data, as the function will not run if the entire column of the spreadsheet is selected. (E.g., If you have 2,841 observations, your selection should only contain 2,842 rows—one for each observation and the title row.)
    a. Be sure to include the column titles in your selection.
8. Make sure the box next to Labels is checked. This tells Excel the first row contains variable names and will make the interpretation of results easier.
9. Select a range for the output data.
    a. The default is to create a new worksheet, or you can select a range within your current worksheet. Either will work, so use whichever you prefer.

10. Click OK to complete the function and perform the regression analysis.

The output should look like this where the dependent variable it risk perceptions of global warming on a 11 point scale and the independent or explanatory variable is ideology on a seven point scale (7 = Strongly Conservative):



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.590170574							
R Square	0.348301306							
Adjusted R Square	0.348041768							
Standard Error	2.479022199							
Observations	2513							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	8247.376032	8247.376	1342.008	9.2575E-236			
Residual	2511	15431.47872	6.145551					
Total	2512	23678.85476						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	10.8186624	0.141889452	76.24712	0	10.54043008	11.09689473	10.54043008	11.09689473
X (ideol)	-1.046346348	0.028562616	-36.6334	9.3E-236	-1.102355044	-0.990337652	-1.102355044	-0.990337652

If you are using Office 365, the steps are slightly different:

1. Select the variables you will be using for your analysis.
   a. If you are performing a multivariate analysis, make sure all independent variables are in columns next to each other, with no gaps or other data in between.
2. Remove any non-valid responses (i.e., “don’t know”, “not applicable”, etc.) from your data before performing any statistical analysis. The easiest way to do this is by sorting each variable and deleting the rows containing non-valid measurements.
   a. Make sure you highlight all columns of data before sorting in order to expand the selection and maintain each observation. If you forget this step, the program will only sort the column you are in and will produce wrong results.
3. On the toolbar at the top of the screen, click the Insert tab.
4. In the middle of the toolbar, select Office Add-Ins.
5. Find XLMiner Analysis Toolpak and click Add to open the toolpak functions on a sidebar.
6. From this menu, select Linear Regression. This should expand the list to give space for inputting variables into the regression model.

7. In the Input Y Range box, highlight your dependent variable. Select only the rows that contain data, as the function will not run if the entire column of the spreadsheet is selected. (E.g., If you have 2,841 observations, your selection should only contain 2,842 rows—one for each observation and the title row.)
   a. It is extremely important that you put the variables into the right boxes at this stage, or the function will return the wrong results. This box should only ever have one variable in it.
   b. Be sure to include the column title in your selection.
8. In the Input X Range box, highlight the columns containing your independent variable(s) and any control variables you are including in your analysis. Select only the rows that contain data, as the function will not run if the entire column of the spreadsheet is selected. (E.g., If you have 2,841 observations, your selection should only contain 2,842 rows—one for each observation and the title row.)
   a. Be sure to include the column titles in your selection.
9. Make sure the box next to Labels is checked. This tells Excel the first row contains variable names and will make the interpretation of results easier.
10. Select a range for the output data by selecting a cell on the current worksheet. (This is different from the standalone version of Excel, in which you can place the regression output on a separate worksheet.)
11. Click OK to complete the function and perform the regression analysis. This should produce the same set of tables generated by the regression function in the standalone version of Excel.

Once you have the output, you can interpret it as follows:

1. The adjusted R-squared for your model is in the third row of the first table of output. This will tell you the percentage of the variation in the DV that is explained by the variable(s) in your model.

   a. Remember there will only be one value of adjusted R-squared for each model, regardless of the number of variables included.

2. The rest of the data you need to look at is in the third table of output. The column titles should include Coefficients, Standard Error, and P-value.

3. The intercept coefficient is the *a* in the regression formula, and the coefficient of the omitted category if using dummy variables.

4. All other variables will be listed in subsequent rows, with their respective coefficients (*b* in the regression formula) in the second column of the table. You can plug these into the formula along with the intercept value to determine the equation of the regression line.

5. Standard errors can be found in the third column of the table (immediately to the right of the coefficients).

6. P-values are located in the fifth column of the same table, which includes a p-value for each variable in the model. These tell you whether each variable has a statistically significant impact on the dependent variable.
    a. If the p-value listed is **less than 0.05**, the relationship IS statistically significant and you CAN reject the null hypothesis for that variable.
    b. If the p-value listed is **greater than 0.05**, the relationship IS NOT statistically significant and you CANNOT reject the null hypothesis for that variable.

Study Questions

Interpret the bivariate regression output coefficients substantively.

OLS regression is the process of minimizing what value? Draw a diagram illustrating this concept. Option Two: Copy and Paste the figure from this chapter and interpret it.

CHAPTER NINE: Bi-Variate Hypothesis Testing and Model Fit

The previous chapters discussed the logic of OLS regression and how to derive OLS estimators. Now that simple regression is no longer a mystery, we will shift the focus to bi-variate hypothesis testing and model fit. We recommend that you try the analyses in the chapter as you read.

Hypothesis Tests for Regression Coefficients

Hypothesis testing is the key to theory building. This chapter is focused on empirical hypothesis testing using OLS regression, with examples drawn from the accompanying class dataset. Here we will use the responses to the political ideology question (ranging from 1=strong liberal, to 7=strong conservative), as well as responses to a question concerning the survey respondents' level of risk that global warming poses for people and the environment.[1]

Using the data from these questions, we posit the following hypothesis:

H1: On average, as respondents become more politically conservative, they will be less likely to express increased risk associated with global warming

The null hypothesis, H0, is $\beta=0$, posits that a respondent's ideology has no relationship with their views about the risks of global warming for people and the environment. Our working hypothesis, H1, is $\beta<0$. We expect β to be less than zero because we expect a *negative* slope between our measures of ideology and levels of risk associated with global warming, given that a larger numeric value for ideology indicates a more conservative respondent. Note that this is a *directional* hypothesis, since we are positing a negative relationship. Typically, a directional hypothesis implies a one-tailed test where the critical value is 0.05 on one side of the distribution. A *non-directional* hypothesis, $\beta\neq0$ does not imply a particular direction, it only implies that there is a relationship. This requires a two-tailed test where the critical value is 0.025 on both sides of the distribution.

The above output tests this hypothesis. So, using our example data, we tested the working hypothesis that political ideology is negatively related to perceived risk of global warming to people and the environment. Using simple OLS regression, we find support for this working hypothesis, and can reject the null.

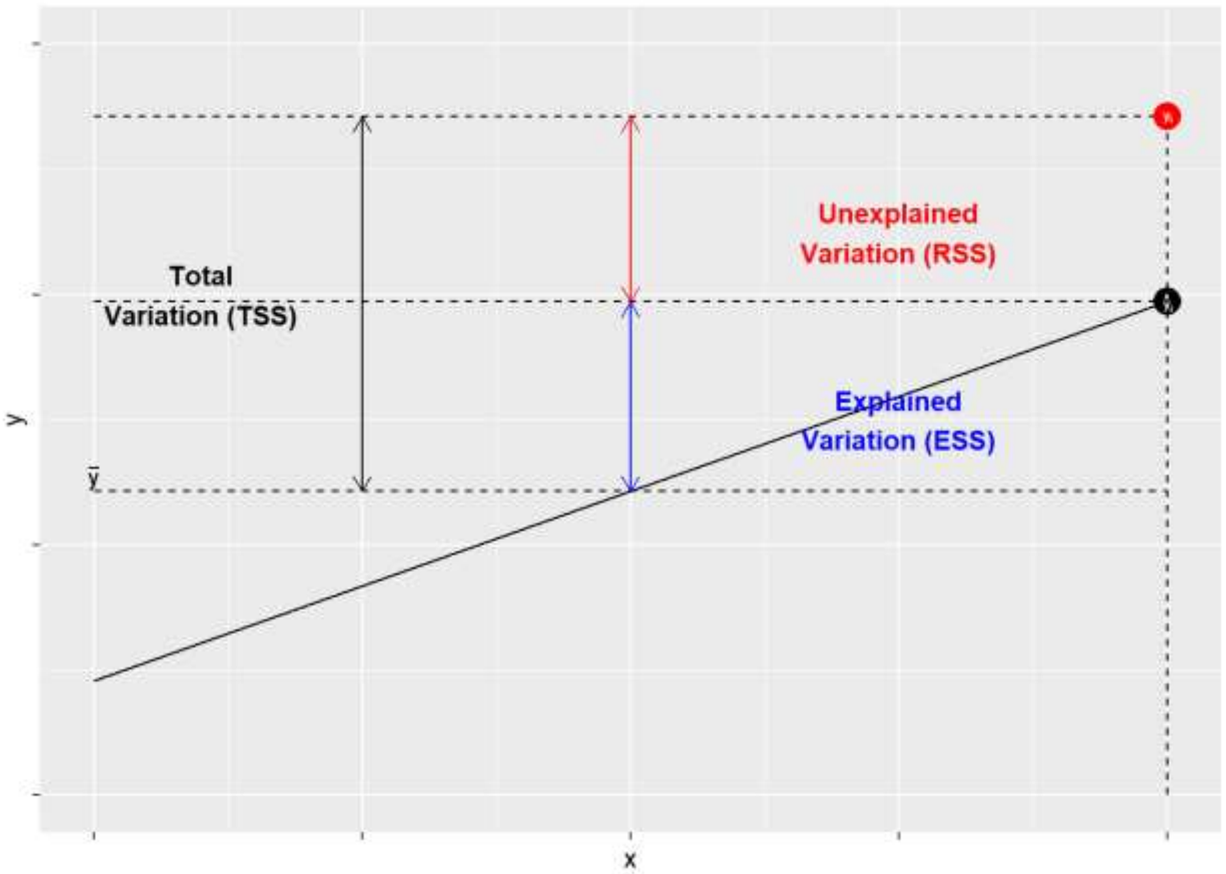
Coefficient of Determination: R^2

The most often used measure of goodness of fit for OLS models is R^2. R^2 is derived from three components: the total sum of squares, the explained sum of squares, and the residual sum of squares. R^2 is the ratio of **ESS** (explained sum of squares) to **TSS** (total sum of squares).

Components of R2R2

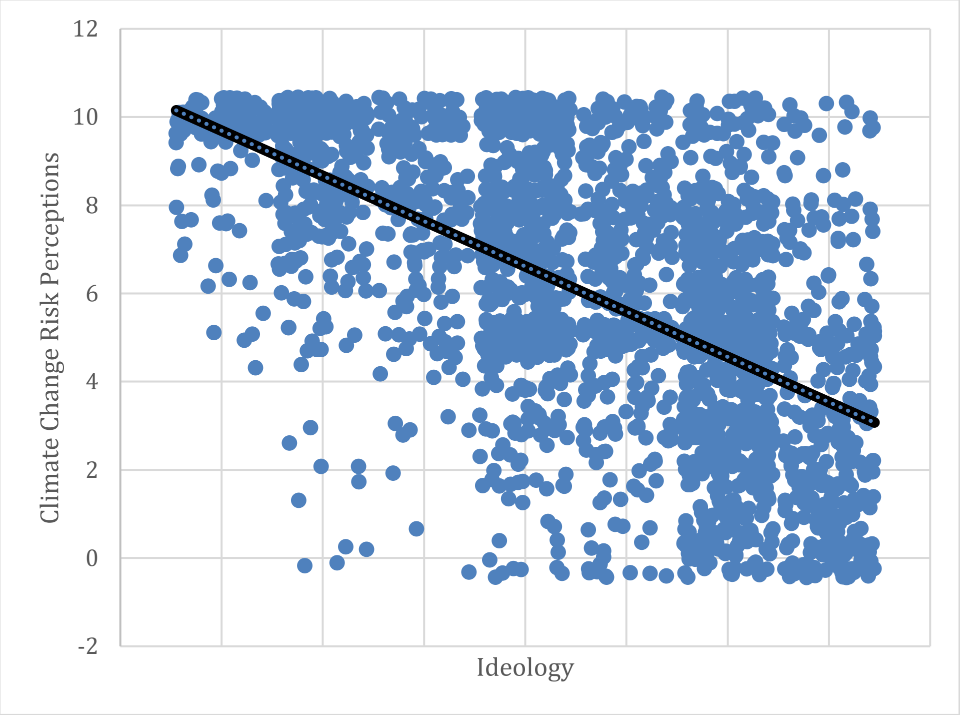
- *Total sum of squares (TSS)*: The sum of the squared variance of Y
- *Residual sum of squares (RSS)*: The variance of Y not accounted for by the model
- *Explained sum of squares (ESS)*: The variance of Y accounted for in the model. It is the difference between the TSS and the RSS.
- R^2: The proportion of the total variance of Y explained by the model, or the ratio of ESS to TSS

The components of R^2 are illustrated in Figure below. As shown, for each observation Y_i, variation around the mean can be decomposed into that which is "explained" by the regression and that which is not. In Figure below the deviation between the mean of Y and the predicted value of Y, $\hat{Y}$, is the proportion of the variation of Y_i that can be explained (or predicted) by the regression. That is shown as a blue line. The deviation of the observed value of Y_i from the predicted value $\hat{Y}$(aka the residual, as discussed in the previous chapter) is the unexplained deviation, shown in red. Together, the explained and unexplained variation make up the total variation of Y_i around the mean , $\hat{Y}$.



Visualizing Bivariate Regression

We have actually already done this when we built the scatterplot before. However, we must note that because the line is an estimate we actually have a realm of uncertainty around it. This would typically be represented by error bars or a “ribbon plot”; however, Excel does not have this functionality built in. This means it is especially important we leave the points in the scatterplot on the graph so that we have some visualization of the uncertainty. Not every single point falls along or even necessarily very close to that line. This is also why the R^2 measure is useful.



Summary

This chapter has focused on two key aspects of simple regression models: hypothesis testing and measures of the goodness of model fit. With respect to the former, we focused on the residual standard error and its role in determining the probability that our model estimates, B and A, are just random departures from a population in which β and α are zero. We showed, using Excel, how to calculate the residual standard errors for A and B and, using them, to calculate the t-statistics and associated probabilities for hypothesis testing. For model fit, we focused on model covariation and correlation, and finished up with a discussion of the coefficient of determination – R^2. So you are now in a position to use simple regression, and to wage unremitting geek-war on those whose models are endowed with lesser R^2s.

Study Questions

1. What is the typical null hypothesis for a regression coefficient? If the p-value is less than 0.05, how do we interpret this coefficient?

2. What is the range of R-squared values? How do we interpret R-squared across this range?

3. What is the interpretation of A (or alpha, also known as the intercept or constant)?

CHAPTER ELEVEN: THE LOGIC OF MULTIPLE REGRESSION

The logic of multiple regression can be readily extended from our earlier discussion of simple regression. As with simple regression, multiple regression finds the regression line (or regression plane" with multiple independent variables) that minimizes the sum of the squared errors. This chapter discusses the theoretical specification of the multiple regression model, the key assumptions necessary for the model to provide the best linear unbiased estimates (BLUE) of the effects of the Xs on Y, the meaning of the partial regression coefficients, and hypothesis testing. Note that the examples in this chapter continue to use the class data set.

Theoretical Specification

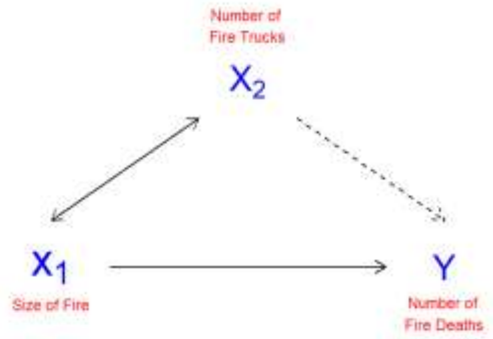
As with simple regression, the theoretical multiple regression model contains a **systematic** component — $Y=\alpha+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_kX_{ik}Y=\alpha+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_kX_{ik}$ and a **stochastic** component—ϵ_i. The overall theoretical model is expressed as:

$$Y=\alpha+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_kX_{ik}+\epsilon_i$$

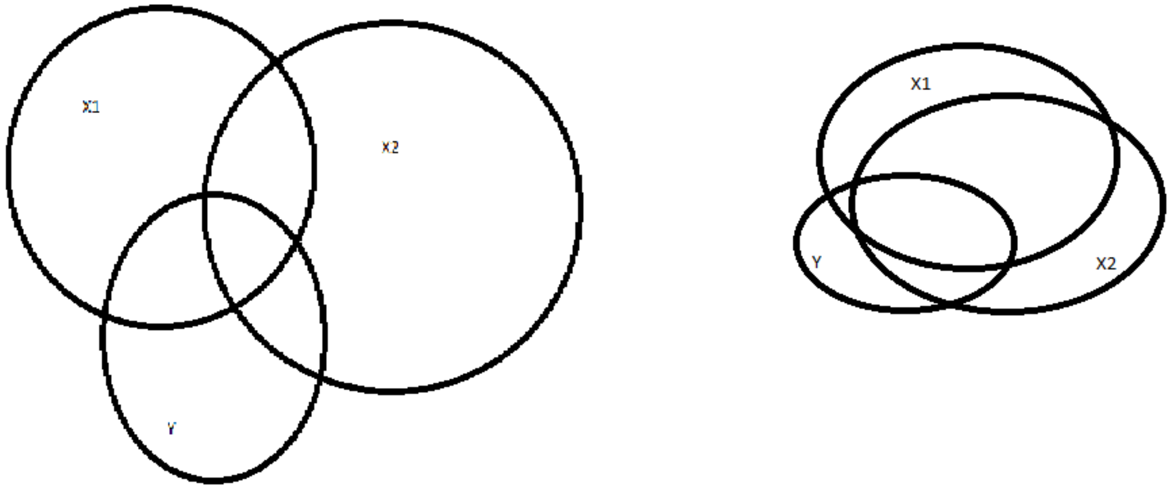
Where α is the constant and each X, denoted by the numeric subscript, is a different independent or explanatory variable and ϵ_i is the error term.

Partial Effects

As noted in Chapter 1, multiple regression controls" for the effects of other variables on the dependent variables. This is in order to manage possible spurious relationships, where the variable Z influences the value of both X and Y. Figure below illustrates the nature of spurious relationships between variables.

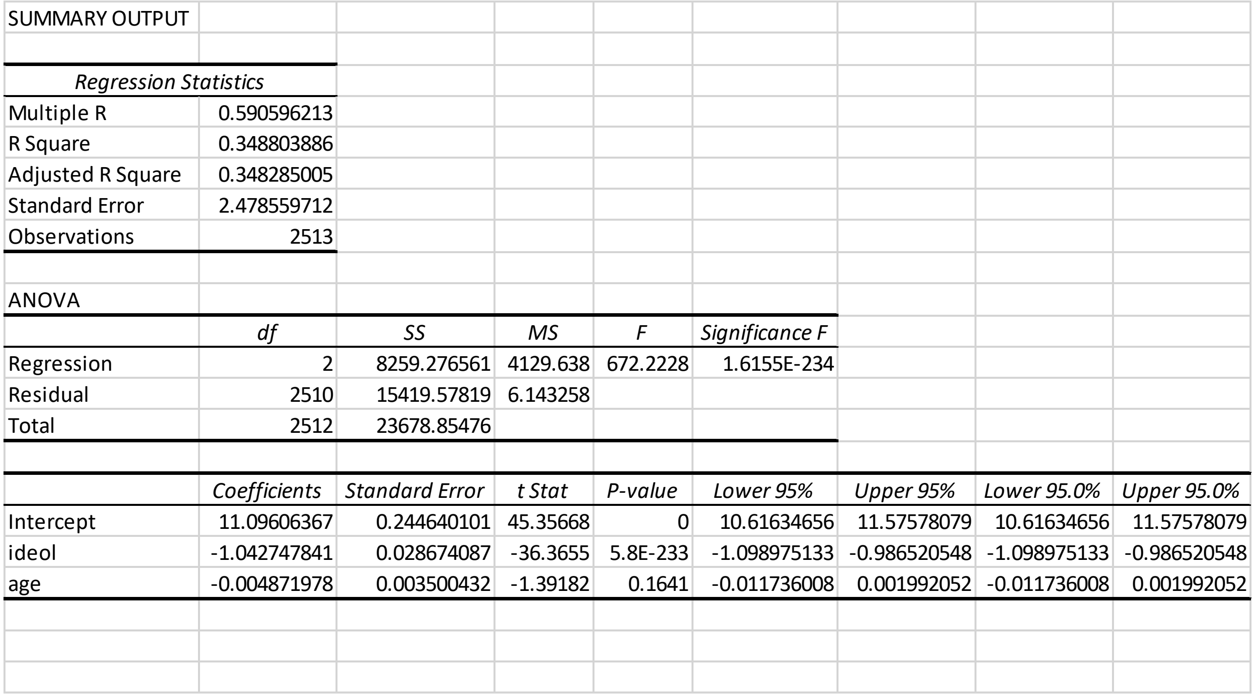


To control for spurious relationships, multiple regression accounts for the **partial effects** of one X on another X. Partial effects deal with the shared variance between Y and the X's. This is illustrated in Figure below. In this example, the number of deaths resulting from house fires is positively associated with the number of fire trucks that are sent to the scene of the fire. A simple-minded analysis would conclude that if fewer trucks are sent, fewer fire-related deaths would occur. Of course, the number of trucks sent to the fire, and the number of fire-related deaths, are both driven by the magnitude of the fire. An appropriate control for the size of the fire would therefore presumably eliminate the positive association between the number of fire trucks at the scene and the number of deaths (and may even reverse the direction of the relationship, as the larger number of trucks may more quickly suppress the fire).



In the figure above, the Venn diagram on the left represents how two variables X1 and X2 can contribute to explaining Y (overlap with Y) and also overlap with each other some. The part a multiple regression will give us an estimate of is the overlap between X1 and Y (for the coefficient on X1) and X2 and Y (for the coefficient on X2) that is unique (meaning not the middle part where all three overlap). This middle part is what is "controlled for" but the estimates that are left are the partial effects. The Venn diagram on the right presents a less optimal estimation because the middle overlap is so large – larger than either of the pairwise overlaps (X1 and Y and X2 and Y).

To estimate multiple regression in Excel, we follow the steps from before but simply add a column to the input X range. Thus, we might be interested in the effect of ideology on climate change risk perceptions after we have accounted for ("controlled for") age. The output for this regression is below:



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.590596213							
R Square	0.348803886							
Adjusted R Square	0.348285005							
Standard Error	2.478559712							
Observations	2513							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	8259.276561	4129.638	672.2228	1.6155E-234			
Residual	2510	15419.57819	6.143258					
Total	2512	23678.85476						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	11.09606367	0.244640101	45.35668	0	10.61634656	11.57578079	10.61634656	11.57578079
ideol	-1.042747841	0.028674087	-36.3655	5.8E-233	-1.098975133	-0.986520548	-1.098975133	-0.986520548
age	-0.004871978	0.003500432	-1.39182	0.1641	-0.011736008	0.001992052	-0.011736008	0.001992052

From this output, we can see that the coefficient on ideology is still negative and significant when controlling for age. If we compare it to the regression output above, the coefficient is ever so slightly less negative going from -1.046 to -1.042 when we account for age. To interpret this, you can say “When accounting for age, the effect of ideology on risk perceptions of climate change is -1.042.” Or “When all else is held constant, the effect of ideology on risk perceptions about global warming is -1.04.” Substantively, you might say, that a one-point increase in ideology, that is becoming more conservative, is associated with a one-point decrease in risk perceptions about global warming.

Summary

The use of multiple regression, when compared to simple bivariate regression, allows for more sophisticated and interesting analyses. The most important feature is the ability of the analyst (that’s you!) to statistically control for the effects of all other IVs when estimating any B. In essence, we clean" the estimated relationship between any X and Y of the influence of all other Xs in the model. Hypothesis testing in multiple regression requires that we identify the independent variation in each X, but otherwise the estimated standard error for each BB is analogous to that for simple regression.

So, maybe it’s a little more complicated. But look at what we can observe! Our estimates from the examples in this chapter show that age, income and education are all related to political ideology, but even when we control for their effects, ideology retains a potent influence on the perceived risks of climate change. Politics matter.

Study Questions

1. Define partial effects.

2. How do we interpret coefficients in multiple regressions? Provide an example.
3. What is the null hypothesis for coefficients in multiple regression?

CHAPTER TWELVE: MULTIPLE REGRESSION AND MODEL BUILDING

This book focuses on the use of systematic quantitative analysis for purposes of building, refining and testing theoretical propositions in the policy and social sciences. All of the tools discussed so far – including univariate, bi-variate, and simple regression analysis – provide means to evaluate distributions and test hypotheses concerning simple relationships. Most policy and social theories, however, include multiple explanatory variables. **Multiple regression** extends the utility of simple regression by permitting the inclusion of two or more explanatory variables. This chapter discusses strategies for determining what variables to include (or exclude) in the model.

Model Building

Model building is the process of deciding which independent variables to include in the model.[1] For our purposes, when deciding which variables to include, theory and findings from the extant literature should be the most prominent guides. Apart from theory, however, this chapter examines empirical strategies that can help determine if the addition of new variables improves overall model fit. In general, when adding a variable, check for: a) improved prediction based on empirical indicators, b) statistically and substantively significant estimated coefficients, and c) stability of model coefficients—do other coefficients change when adding the new one – particularly look for sign changes.

Theory and Hypotheses

The most important guidance for deciding whether a variable (or variables) should be included in your model is provided by theory and prior research. Simply put, knowing the literature on your topic is vital to knowing what variables are important. You should be able to articulate a clear theoretical reason for including each variable in your model. In those cases where you don't have much theoretical guidance, however, you should use model *parsimony*, which is a function of simplicity and model fit, as your guide. You can focus on whether the inclusion of a variable improves model fit. In the next section, we will explore several empirical indicators that can be used to evaluate the appropriateness of variable inclusion.

Empirical Indicators

When building a model, it is best to start with a few IV's and then begin adding other variables. However, when adding a variable, check for:

- Improved prediction (increase in adjusted R^2)
- Statistically and substantively significant estimated coefficients

- Stability of model coefficients
  - Do other coefficients change when adding the new one?
  - Particularly look for sign changes for estimated coefficients.

Coefficient of Determination: R^2

R^2 was previously discussed within the context of simple regression. The extension to multiple regression is straightforward, except that multiple regression leads us to place greater weight on the use of the **adjusted** R^2. Recall that the adjusted R^2 corrects for the inclusion of multiple independent variables; R^2 is the ratio of the explained sum of squares to the total sum of squares (*ESS/TSS*).

R^2 is expressed as:

$$R^2 = 1 - \frac{RSS}{TSS}$$

However, this formulation of R^2 is insensitive to the complexity of the model and the degrees of freedom provided by your data. This means that an increase in the number of kk independent variables, can increase the R^2. Adjusted R^2 penalizes the R^2 by correcting for the degrees of freedom. It is defined as:

$$adjusted\ R^2 = 1 - \frac{\frac{RSS}{n-k-1}}{\frac{TSS}{n-k-1}}$$

The R^2 of two models can be compared, as illustrated by the following example. The first (simpler) model consists of basic demographics (age, education, and income) as predictors of climate change risk. The second (more complex) model adds the variable measuring political ideology to the explanation.

As can be seen by comparing the model results, the more complex model that includes political ideology has a higher R^2 than does the simpler model. This indicates that the more complex model explains a greater fraction of the variance in perceived risks of climate change. However, we don't know if this improvement is statistically significant. In order to determine whether the more complex model adds significantly to the explanation of perceive risks, we can utilize the F-test.

Risks in Model Building

As is true of most things in life, there are risks to consider when building statistical models. First, are you including irrelevant X's? These can increase model complexity, reduce adjusted R^1, and increase model variability across samples. Remember that you should have a theoretical basis for inclusion of all of the variables in your model.

Second, are you omitting relevant X's? Not including important variables can fail to capture fit and can bias other estimated coefficients, particularly when the omitted X is related to both other X's and to the dependent variable Y.

Finally, remember that we are using sample data. Therefore, about 5% of the time, our sample will include random observations of X's that result in B's that meet classical hypothesis tests – resulting in a Type I error. Conversely, the B's may be important, but the sample data will randomly include observations of X that result in estimated parameters that do not meet the classical statistical tests – resulting in a Type II error. That's why we rely on theory, prior hypotheses, and replication.

Evils of Stepwise Regression

Almost all statistical software packages permit a number of mechanical "search strategies" for finding IVs that make a statistically significant contribution to the prediction of the model dependent variable. The most common of these is called **stepwise regression**, which may also be referred to as forward, backward (or maybe even upside down!) stepwise regression. Stepwise procedures do not require that the analyst think – you just have to designate a pool of possible IVs and let the package go to work, sifting through the IVs to identify those that (on the basis of your sample data) appear to be related to the model dependent variable. The stepwise procedures use sequential F-tests, sequentially adding variables that "improve the fit" of the mindless model until there are no more IVs that meet some threshold (usually p<0.05) of statistical significance. These procedures are like mechanically wringing all of the explanation you can get for Y out of some pool ofXX.

You should already recognize that these kind of methods pose serious problems. First and foremost, this is an atheoretical approach to model building. But, what if you have no theory to start with – is a stepwise approach appropriate then? No, for several reasons. If any of the candidate X variables are strongly correlated, the inclusion of the first one will "use up" some of the explanation of the second, because of the way OLS calculates partial regression coefficients. For that reason, once one of the variables is mechanically selected, the other will tend to be excluded because it will have less to contribute to Y. Perhaps more damning, stepwise approaches are highly susceptible to inclusion of spuriously related variables. Recall that we are using samples, drawn from the larger population, and that samples are subject to random variation. If the step-wise process uses the classical 0.05 cut-off for inclusion of a variable, that means that one time in twenty (in the long run) we will include a variable that meets the criterion only by random chance.[2] Recall that the classical hypothesis test requires that we specify our hypothesis in advance; step-wise processes simply rummage around within a set of potential IVs to find those that fit.

There have been notable cases in which mechanical model building has resulted in seriously problematic "findings" that have very costly implications for society. One is recounted in the PBS Frontline episode called "Currents of Fear".^[The program was written, produced and directed by Jon Palfreman, and it was first broadcast on June 13, 1995. The full transcript can be found here. The story concerns whether electromagnetic fields (EMFs) from technologies including high-voltage power lines cause cancer in people who are exposed. The problem was that "cancer clusters" could be identified that were proximate to the power lines, but no laboratory experiments could find a connection. However, concerned citizens and activists persisted in believing there was a causal relationship. In that context, the Swedish government sponsored a very ambitious study to settle the question. Here is the text of the discussion from the Frontline program:

… in 1992, a landmark study appeared from Sweden. A huge investigation, it enrolled everyone living within 300 meters of Sweden's high-voltage transmission line system over a 25-year period. They went far beyond all previous studies in their efforts to measure magnetic fields, calculating the fields that the children were exposed to at the time of their cancer diagnosis and before. This study reported an apparently clear association between magnetic field exposure and childhood leukemia, with a risk ratio for the most highly exposed of nearly 4.

The Swedish government announced it was investigating new policy options, including whether to move children away from schools near power lines. Surely, here was the proof that power lines were dangerous, the proof that even the physicists and biological naysayers would have to accept. But three years after the study was published, the Swedish research no longer looks so unassailable. This is a copy of the original contractor's report, which reveals the remarkable thoroughness of the Swedish team. Unlike the published article, which just summarizes part of the data, the report shows everything they did in great detail, all the things they measured and all the comparisons they made.

When scientists saw how many things they had measured – nearly 800 risk ratios are in the report – they began accusing the Swedes of falling into one of the most fundamental errors in epidemiology, sometimes called the multiple comparisons fallacy.

So, according to the Frontline report, the Swedish EMF study regressed the incidence of nearly 800 possible cancers onto the proximity of its citizens to high-voltage power lines. In some cases, there appeared to be a positive relationship. These they reported. In other cases, there was no relationship, and in some the relationship was negative - which would seem to imply (if you were so silly as to do so) that living near the high voltage lines actually protected people from cancer. But only the positive relationships were included in the reports, leading to a false impression that the study had confirmed that proximity to high-voltage lines causes cancer. Embarrassing to the study authors, to put it mildly.

Summary

This chapter has focused on multiple regression model building. The keys to that process are understanding (a) the critical role of theory and prior research findings in model specification, and (b) the meaning of the partial regression coefficients produced by OLS. When theory is not well-developed, you can thoughtfully employ nested F-tests to evaluate whether the hypothesized inclusion of an X variable meaningfully contributes to the explanation of Y. But you should avoid reliance on mechanical model-building routines, like step-wise regression, because these can lead you down into statistical perdition. None of us want to see that happen!

Study Questions

1. Why is *adjusted* R-squared a better measure of goodness of fit than regular R-squared in multiple regression?

2. How can we use fit statistics to help use build and assess out theoretical model?

CHAPTER THIRTEEN: Topic in Multiple Regression

Thus far we have developed the basis for multiple OLS regression using matrix algebra, delved into the meaning of the estimated partial regression coefficient, and revisited the basis for hypothesis testing in OLS. In this chapter we turn to one of the key strengths of OLS: the robust flexibility of OLS for model specification. First we will discuss how to include binary variables (referred to as dummy variables") as IVs in an OLS model. Next we will show you how to build on dummy variables to model their interactions with other variables in your model. Finally, we will address an alternative way to express the partial regression coefficients – using standardized coefficients – that permit you to compare the magnitudes of the estimated effects of your IVs even when they are measured on different scales. As has been our custom, the examples in this chapter are based on variables from the class data set.

Dummy Variables

Thus far, we have considered OLS models that include variables measured on interval level scales (or, in a pinch and with caution, ordinal scales). That is fine when we have variables for which we can develop valid and reliable interval (or ordinal) measures. But in the policy and social science worlds, we often want to include in our analysis concepts that do not readily admit to interval measure – including many cases in which a variable has an "on - off", or "present - absent" quality. In other cases we want to include a concept that is essentially nominal in nature, such that an observation can be categorized as a subset but not measured on a "high-low" or "more-less" type of scale. In these instances we can utilize what is generally known as a dummy variable, but are also referred to as indicator variables, Boolean variables, or categorical variables.

What the Heck are "Dummy Variables"?

- A dichotomous variable, with values of 0 and 1;

- A value of 1 represents the presence of some quality, a zero its absence;

- The 1s are compared to the 0s, who are known as the referent group";

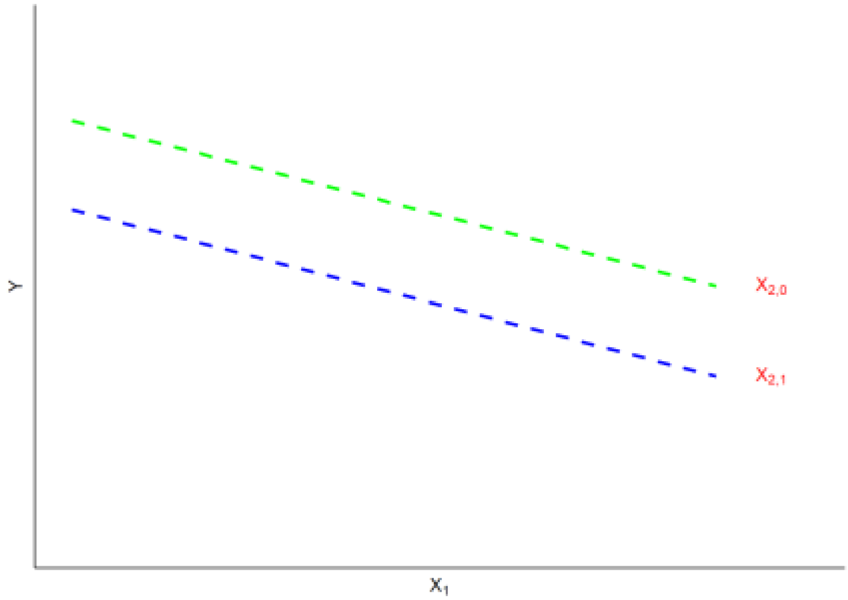
- Dummy variables are often thought of as a proxy for a qualitative variable.

Dummy variables allow for tests of the differences in overall value of the YY for different nominal groups in the data. They are akin to a difference of means test for the groups identified by the dummy variable. Dummy variables allow for comparisons between an included (the 1s) and an omitted (the 0s) group. Therefore, it is important to be clear about which group is omitted and serving as the comparison category."

It is often the case that there are more than two groups represented by a set of nominal categories. In that case, the variable will consist of two or more dummy variables, with 0/1 codes for each category except the referent group (which is omitted). Several examples of categorical variables that can be represented in multiple regression with dummy variables include:

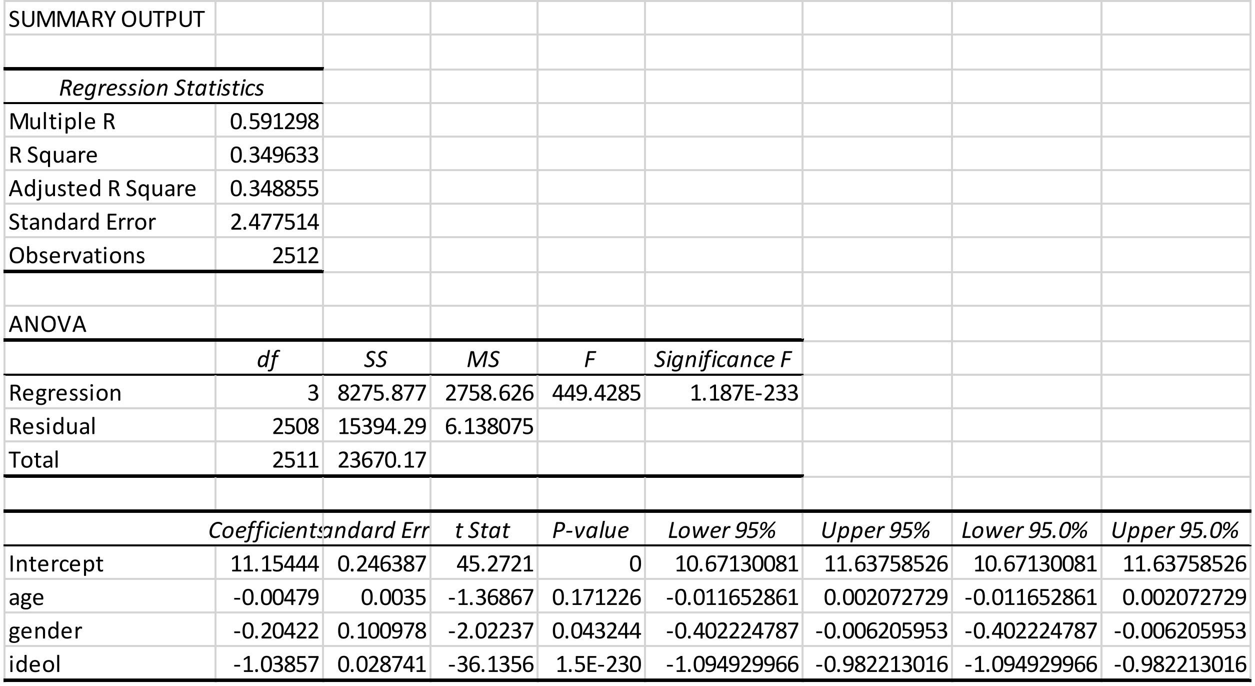
- Experimental treatment and control groups (treatment=1, control=0)
- Gender (male=1, female=0 or vice versa)
- Race and ethnicity (a dummy for each group, with one omitted referent group)
- Region of residence (dummy for each region with one omitted reference region)
- Type of education (dummy for each type with omitted reference type)
- Religious affiliation (dummy for each religious denomination with omitted reference)

The value of the dummy coefficient represents the estimated difference in Y between the dummy group and the reference group. Because the estimated difference is the average over all of the Y observations, the dummy is best understood as a change in the value of the intercept (A) for the dummied" group. This is illustrated in following figure. In this illustration, the value of YY is a function of X1 (a continuous variable) and X2 (a dummy variable). When X2 is equal to 0 (the referent case) the top regression line applies. When X2=1, the value of Y is reduced to the bottom line. In short, X2 has a negative estimated partial regression coefficient represented by the difference in height between the two regression lines.



For a case with multiple nominal categories (e.g., region) the procedure is as follows: (a) determine which category will be assigned as the referent group; (b) create a dummy variable for each of the other categories. For example, if you are coding a dummy for four regions (North, South, East and West), you could designate the South as the referent group. Then you would create dummies for the other three regions. Then, all observations from the North would get a value of 1 in the North dummy, and zeros in all others. Similarly, East and West observations would receive a 1 in their respective dummy category and zeros elsewhere. The observations from the South region would be given values of zero in all three categories. The interpretation of the partial regression coefficients for each of the three dummies would then be the estimated difference in Y between observations from the North, East and West and those from the South.

Now let's walk through an example of a regression model with a dummy variable and the interpretation of that model. We will predict climate change risk using age, income, ideology, and "gend", a dummy variable for gender for which 1 = male and 0 = female.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.591298
R Square	0.349633
Adjusted R Square	0.348855
Standard Error	2.477514
Observations	2512

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	8275.877	2758.626	449.4285	1.187E-233
Residual	2508	15394.29	6.138075		
Total	2511	23670.17			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	11.15444	0.246387	45.2721	0	10.67130081	11.63758526	10.67130081	11.63758526
age	-0.00479	0.0035	-1.36867	0.171226	-0.011652861	0.002072729	-0.011652861	0.002072729
gender	-0.20422	0.100978	-2.02237	0.043244	-0.402224787	-0.006205953	-0.402224787	-0.006205953
ideol	-1.03857	0.028741	-36.1356	1.5E-230	-1.094929966	-0.982213016	-1.094929966	-0.982213016

In this case, the interpretation of the coefficients on age and ideology are the same as they would be above. The interpretation of gender is different however. The coefficient on gender reflects the difference in global warming risk perceptions for males, relative to females. First note that the inclusion of the dummy variables does not change the manner in which you interpret the other (non-dummy) variables in the model; the estimated partial regression coefficients for age, education, income and ideology should all be interpreted as described in the prior chapter. Note that the estimated partial regression coefficient for gender" is negative and statistically significant, indicating that males are less likely to be concerned about the environment than are females. The estimate indicates that, all else being equal, the average difference between men and women on the climate change risk scale is -0.204.

Summary

This chapter has focused on options in designing and using OLS models. We covered the use of dummy variables to capture the effects of group differences on estimates of Y. Overall, these refinements in the use of OLS permit great flexibility in the application of regression models to estimation and hypothesis testing in policy analysis and social science research.

Study Questions

1. What is a dummy variable? When should we use it? How do you interpret coefficients on dummy variables?