Statistics Minus The Math: An Introduction for the Social Sciences

Nathan Favero

2024-02-11

Table of contents

Introduction

This version (1.2) was updated 1/28/2023.

Version 1.1 available at https://nathanfavero.com/teaching-tools/. The only change from 1.1 is that the discussion of transforming variables now appears in Ch. 2 (rather than Ch. 3).

This book was largely adapted from the public domain resource *Online Statistics Education: A Multimedia Course of Study* (http://onlinestatbook.com/ Project Leader: David M. Lane, Rice University).

1 Causality

1.1 Causation[1]

The concept of causation is a complex one in the philosophy of science.[2] Since a full coverage of this topic is well beyond the scope of this text, we focus on two specific topics: (1) the establishment of causation in experiments and (2) the establishment of causation in non-experimental designs.

1.1.1 Establishing Causation in Experiments

Consider a simple experiment in which subjects are sampled randomly from a population and then assigned randomly to either the experimental group or the control group. Assume the condition means on the dependent variable differed. Does this mean the treatment caused the difference?

To make this discussion more concrete, assume that the experimental group received a drug for insomnia, the control group received a placebo, and the dependent variable was the number of minutes the subject slept that night. An obvious obstacle to inferring causality is that there are many unmeasured variables that affect how many hours someone sleeps. Among them are how much stress the person is under, physiological and genetic factors, how much caffeine they consumed, how much sleep they got the night before, etc. Perhaps differences between the groups on these factors are responsible for the difference in the number of minutes slept.

At first blush it might seem that the random assignment eliminates differences in unmeasured variables. However, this is not the case. Random assignment ensures that differences on unmeasured variables are chance differences. It does not ensure that there are no differences. Perhaps, by chance, many subjects in the control group were under high stress and this stress made it more difficult to fall asleep. The fact that the greater stress in the control group was due to chance does not mean it could not be responsible for the difference between the control and the experimental groups. In other words, the observed difference in "minutes slept" could have been due to a chance difference between the control group and the experimental group rather than due to the drug's effect.

[1] This section is adapted from David M. Lane. "Causation." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/research_design/causation.html

[2] See http://plato.stanford.edu/search/searcher.py?query=causation

This problem seems intractable since, by definition, it is impossible to measure an "unmeasured variable" just as it is impossible to measure and control all variables that affect the dependent variable. However, although it is impossible to assess the effect of any single unmeasured variable, it is possible to assess the combined effects of all unmeasured variables. Since everyone in a given condition is treated the same in the experiment, differences in their scores on the dependent variable must be due to the unmeasured variables. Therefore, a measure of the differences among the subjects within a condition is a measure of the sum total of the effects of the unmeasured variables. The most common measure of differences is the variance. By using the within-condition variance to assess the effects of unmeasured variables, statistical methods determine the probability that these unmeasured variables could produce a difference between conditions as large or larger than the differenc[3]e obtained in the experiment. If that probability is low, then it is inferred (that's why they call it inferential statistics) that the treatment had an effect and that the differences are not entirely due to chance. Of course, there is always some nonzero probability that the difference occurred by chance so total certainty is not a possibility.

1.1.2 Causation in Non-Experimental Designs

It is almost a cliché that correlation does not mean causation. The main fallacy in inferring causation from correlation is called the third variable problem and means that a third variable is responsible for the correlation between two other variables. An excellent example used by Li (1975) to illustrate this point is the positive correlation in Taiwan in the 1970's between the use of contraception and the number of electric appliances in one's house. Of course, using contraception does not induce you to buy electrical appliances or vice versa. Instead, the third variable of education level affects both.

Does the possibility of a third-variable problem make it impossible to draw causal inferences without doing an experiment? One approach is to simply assume that you do not have a third-variable problem. This approach, although common, is not very satisfactory. However, be aware that the assumption of no third-variable problem may be hidden behind a complex causal model that contains sophisticated and elegant mathematics.

A better though, admittedly more difficult approach, is to find converging evidence. This was the approach taken to conclude that smoking causes cancer. The analysis included converging evidence from retrospective studies, prospective studies, lab studies with animals, and theoretical understandings of cancer causes.

A second problem is determining the direction of causality. A correlation between two variables does not indicate which variable is causing which. For example, Reinhart and Rogoff (2010)[4] found a strong correlation between public debt and GDP growth. Although some have

[3]Li, C. (1975) *Path analysis: A primer.* Boxwood Press, Pacific Grove, CA.

[4]Reinhart, C. M. and Rogoff, K. S. (2010). Growth in a Time of Debt. Working Paper 15639, National Bureau of Economic Research, http://www.nber.org/papers/w15639

argued that public debt slows growth, most evidence supports the alternative that slow growth increases public debt.[5]

1.2 Experimental Designs[6]

There are many ways an experiment can be designed. For example, subjects can all be tested under each of the treatment conditions or a different group of subjects can be used for each treatment. An experiment might have just one independent variable or it might have several. This section describes basic experimental designs and their advantages and disadvantages.

1.2.1 Between-Subjects Designs

In a **between-subjects** design, the various experimental treatments are given to different groups of subjects. For example, in the "Teacher Ratings"[7] case study, subjects were randomly divided into two groups. Subjects were all told they were going to see a video of an instructor's lecture after which they would rate the quality of the lecture. The groups differed in that the subjects in one group were told that prior teaching evaluations indicated that the instructor was charismatic whereas subjects in the other group were told that the evaluations indicated the instructor was punitive. In this experiment, the independent variable is "Condition" and has two levels (charismatic teacher and punitive teacher). It is a between-subjects variable because different subjects were used for the two levels of the independent variable: subjects were in either the "charismatic teacher" or the "punitive teacher" condition. Thus the comparison of the charismatic-teacher condition with the punitive-teacher condition is a comparison between the subjects in one condition with the subjects in the other condition.

The two conditions were treated exactly the same except for the instructions they received. Therefore, it would appear that any difference between conditions should be attributed to the treatments themselves. However, this ignores the possibility of chance differences between the groups. That is, by chance, the raters in one condition might have, on average, been more lenient than the raters in the other condition. Randomly assigning subjects to treatments ensures that all differences between conditions are chance differences; it does not ensure there will be no differences. The key question, then, is how to distinguish real differences from chance differences. The field of inferential statistics answers just this question. Analyzing the data from this experiment reveals that the ratings in the charismatic-teacher condition were higher than those in the punitive-teacher condition. Using inferential statistics, it can be calculated that the probability of finding a difference as large or larger than the one obtained

[5] For a video on causality featuring evidence that smoking causes cancer, see http://www.learner.org/resources/series65.html

[6] This section is adapted from David M. Lane. "Experimental Designs." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/research_design/designs.html

[7] http://onlinestatbook.com/2/case_studies/ratings.html

if the treatment had no effect is only 0.018. Therefore it seems likely that the treatment had an effect and it is not the case that all differences were chance differences.

Independent variables often have several levels. For example, in the "Smiles and Leniency" case study the independent variable is "type of smile" and there are four levels of this independent variable: (1) false smile, (2) felt smile, (3) miserable smile, and (4) a neutral control. Keep in mind that although there are four levels, there is only one independent variable. Designs with more than one independent variable are considered next.

1.2.2 Multi-Factor Between-Subject Designs

In the "Bias Against Associates of the Obese"[8] experiment, the qualifications of potential job applicants were judged. Each applicant was accompanied by an associate. The experiment had two independent variables: the weight of the associate (obese or average) and the applicant's relationship to the associate (girl friend or acquaintance). This design can be described as an Associate's Weight (2) x Associate's Relationship (2) factorial design. The numbers in parentheses represent the number of levels of the independent variable. The design was a factorial design because all four combinations of associate's weight and associate's relationship were included. The dependent variable was a rating of the applicant's qualifications (on a 9-point scale).

If two separate experiments had been conducted, one to test the effect of Associate's Weight and one to test the effect of Associate's Relationship then there would be no way to assess whether the effect of Associate's Weight depended on the Associate's Relationship. One might imagine that the Associate's Weight would have a larger effect if the associate were a girl friend rather than merely an acquaintance. A factorial design allows this question to be addressed. When the effect of one variable does differ depending on the level of the other variable then it is said that there is an interaction between the variables.

Factorial designs can have three or more independent variables. In order to be a between-subjects design there must be a separate group of subjects for each combination of the levels of the independent variables.

1.2.3 Within-Subjects Designs

A **within-subjects** design differs from a between-subjects design in that the same subjects perform at all levels of the independent variable. For example consider the "ADHD Treatment"[9] case study. In this experiment, subjects diagnosed as having attention deficit disorder were each tested on a delay of gratification task after receiving methylphenidate (MPH). All

[8] http://onlinestatbook.com/2/case_studies/obesity_relation.html
[9] http://onlinestatbook.com/2/case_studies/adhd.html

subjects were tested four times, once after receiving one of the four doses. Since each subject was tested under *each* of the four levels of the independent variable "dose," the design is a within-subjects design and dose is a within-subjects variable. Within-subjects designs are sometimes called repeated-measures designs.

1.2.4 Advantage of Within-Subjects Designs

An advantage of within-subjects designs is that individual differences in subjects' overall levels of performance are controlled. This is important because subjects invariably will differ greatly from one another. In an experiment on problem solving, some subjects will be better than others regardless of the condition they are in. Similarly, in a study of blood pressure some subjects will have higher blood pressure than others regardless of the condition. Within-subjects designs control these individual differences by comparing the scores of a subject in one condition to the scores of the same subject in other conditions. In this sense each subject serves as his or her own control. This typically gives within-subjects designs considerably more power (ability to find precise estimates) than between-subjects designs. That is, this makes within-subjects designs more able to detect an effect of the independent variable than are between-subjects designs.

Within-subjects designs are often called "repeated-measures" designs since repeated measurements are taken for each subject. Similarly, a within-subject variable can be called a repeated-measures factor.

1.2.5 Complex Designs

Designs can contain combinations of between-subject and within-subject variables. For example, the "Weapons and Aggression"[10] case study has one between-subject variable (gender) and two within-subject variables (the type of priming word and the type of word to be responded to).

[10] http://onlinestatbook.com/2/case_studies/guns.html

2 Models and Uncertainty

Before I leave my house each morning, I need to decide whether to take an umbrella. So I check my phone to see whether it's supposed to rain. Instead of giving me a direct yes or no answer, the weather tells me the percentage chance of rain for the day.

Why does the weather app give me a percentage? Because there's uncertainty. Science has done a lot to help us understand the weather. And as our understanding of the weather improves, our predictions get better. But we still can't predict rain perfectly.

Facing uncertainty is a common problem when we're looking at data. Whether we're trying to explain the weather, human behavior, or even plant growth, we can't make perfect predictions because there are things we can't fully explain with our current scientific knowledge.

In statistics, we have several tools that allow us to acknowledge uncertainty. This enables us to build models like the ones powering my weather app—models that give us a prediction that includes a description of how uncertain we are. Some days we are 100% sure it will rain, other days only 60%.

In order to build these models that acknowledge uncertainty, we need a way to talk about what we do know and what we don't know. Let me give a very simple example of a model that accounts for uncertainty:

$$happiness = 3.0 + 2.3 \times income + \varepsilon \tag{2.1}$$

This model attempts to explain one's level of happiness based on their income. You might notice that it looks very similar to the regression equations we saw in Chapter 3. That's because regression is one of the main tools used to estimate a model that includes uncertainty.

What does this model mean in practical terms? Well, there are no obvious units we can use to quantify the amount of happiness someone experiences, so the exact values of the numbers we see are not particularly meaningful. But the fact that there's a positive number (2.3) that is being multiplied by income implies that as income gets bigger, happiness gets larger.

The key part of this equation that I want to focus on is the little Greek letter at the end of the equation: ε. This letter is called "epsilon," and it is often used to represent what we call an **error term** (also sometimes called a **disturbance term**). The error term (ε) represents everything else besides income that affects happiness. By including an error term, we are acknowledging that we can't perfectly predict one's level of happiness based on their income.

We think that knowing one's income will help us predict their happiness, but we know there are other factors we won't be able to measure or identify that will also affect happiness. Thus, if all we know about someone is their income, we will have uncertainty about their exact level of happiness. By including an error term (ε) in the model, we make clear that we only claim to have a partial understanding of happiness, not a complete one.

Think for a moment about how few topics we could study if we didn't have the freedom to build models that include uncertainty. We'd only be able to build a model of a dependent variable after we had identified (and measured) *all* of the factors that affect that variable! We wouldn't be able to build a model of rain since we don't know all of the factors that affect the rain. We couldn't build a model of voting behavior since we don't know everything that affects how someone will vote. By including an error term in our model, we can build models even when our understanding of something is incomplete.

The first part of our model that appears on the right side of the equation ($3.0 + 2.3 \times income$) is sometimes described as the *systematic* part of our model. It's what we would use to build a prediction of happiness if all we know about some is their income level. Suppose, for example, that someone has an income of 4 units (perhaps income is measured in tens of thousands of dollars of annual income, so a salary of $40,000 is coded as a 4). According to our model, that person's happiness would be:

$$happiness = 3.0 + 2.3 \times (4) + \varepsilon$$

$$happiness = 12.2 + \varepsilon$$

We, therefore, predict that someone with an income of 4 will have a happiness of 12.2, but we also acknowledge that their actual happiness will likely be a bit different from our prediction since our model indicates that their actual happiness will equal 12.2 plus the value of the error term (ε).

The error term describes something unknown, so we can't measure it or directly observe it. But what we can do is talk about its characteristics using concepts from probability theory. Specifically, we're going to describe the value of the error term as being randomly selected. You may have dealt with randomness in math classes before using examples such as coin flips, die rolls, or drawing cards from a 52-card deck. Just as the likelihood of different outcomes from parlor games can be described using probability, we're going to use probability to describe different possible values for the error term of a statistical model.

2.1 Assumptions about error terms

It's easy to write out an equation that includes an error term, but we are not going to be able to do much with our model unless we make some assumptions about the error term. One of the most important (and challenging) parts of doing statistical analysis is making assumptions

about the possible values of the error term. Different assumptions about the error term can result in very different conclusions.

Let's again consider the simple model of happiness that was introduced above:

$$happiness = 3.0 + 2.3 \times income + \varepsilon$$

We might assume the following things about the error term (ε):

1. The values of the error term (ε) can be described by a normal distribution with a mean of 0
2. Knowing someone's income doesn't help us predict the values of the error term (ε)

What do these two assumptions mean?

First, if the error term (ε) follows a normal distribution with a mean of zero, that means that (according to our model), people are just as likely to have a positive value of the error term as they are to have a negative value of the error term. In other words, all those factors we haven't accounted for in our model are equally likely to push people in the direction of being happier or in the direction of being less happy. Our model and assumptions tell us that if we predict happiness purely based on income, we'll *overestimate* some people's happiness, and we'll *underestimate* an equal number of people's happiness.[1]

Second, these assumptions allow us to describe how much individual observations will tend to deviate from our income-based predictions. We haven't specified in our assumptions what the standard deviation is for the normal distribution for the error term (ε), but statistical analysis will let us estimate the standard deviation of an error term. And we know that there is a 95% chance of drawing a value within two standard deviations of the mean for any normal distribution. So whatever the standard deviation of the error term (ε) is, we would expect that 95% of the time, the error term will take on a value that is within two standard deviations of zero. Conversely, 5% of the time, the error term will take on a value that is more than two standard deviations from zero. Suppose that the standard deviation of the error term (ε) happens to be three. If we have a dataset containing the income and happiness of 1,000 randomly selected people, we would expect that about 950 of these people will have a level of happiness that falls within six units of our income-based prediction. But for about 50 of these people, our prediction of their happiness will be off by more than six units.

Third, our assumptions imply that income is not tied in any consistent way to (the total sum of) factors other than income that also affect peoples' happiness. Remember, the error term (ε) represents all factors other than income that affect satisfaction. If income is related to these other factors, then the value of income should help us predict the value of the error term. For example, if having a stable environment in childhood tends to cause both higher incomes

[1] Note that these deviations from our prediction don't imply that our model is wrong; our model explicitly acknowledges that we'll get only imperfect estimates if we predict happiness based on income, since the unobserved error term (ε) also contributes to happiness.

and greater happiness in adulthood, the error term will partially reflect the effect of childhood stability on happiness, so high incomes (which are partially caused by childhood stability) will be probably be predictive of a more positive error term. This would constitute a violation of our assumptions since we specifically indicated that income wasn't predictive of the error term. As this example illustrates, our assumptions about error terms are often quite strict, making it rather difficult in practice to build good models that account for uncertainty.

2.2 Models and probabilistic thinking

Despite the difficulty inherent in building models that accommodate uncertainty, we have little alternative unless we wish to only build models of things we think we can predict with 100% accuracy. And fortunately, our models do not always have to be perfectly correct in order to generate useful predictions or explanations. As the statistician George Box famously said, "all models are wrong, but some are useful."

An important part of learning to do good statistical analysis is learning to think clearly about models so that you can pick out a model that is useful for whatever it is you want to accomplish. And the first step toward understanding many statistical models is learning to think about the world in probabilistic terms, as we've done here in this reading. Probabilistic thinking asks questions like:

- Based on what I do know and what I don't know, what can I predict?
- How does adding or removing different pieces of information change my prediction?
- How much uncertainty is there in my prediction?
- How often will my prediction differ greatly from what actually happens (even if my model is correct)?

3 Regression with Qualitative Independent Variables

Let's say I'm interested in studying how personality relates to gender. The most common personality measure in psychology is called the "Big Five" personality inventory. There is a standard set of 50 survey items that researchers can use to measure five aspects of personality. Figure 3.1 is an example of some of these questions and how they are formatted:

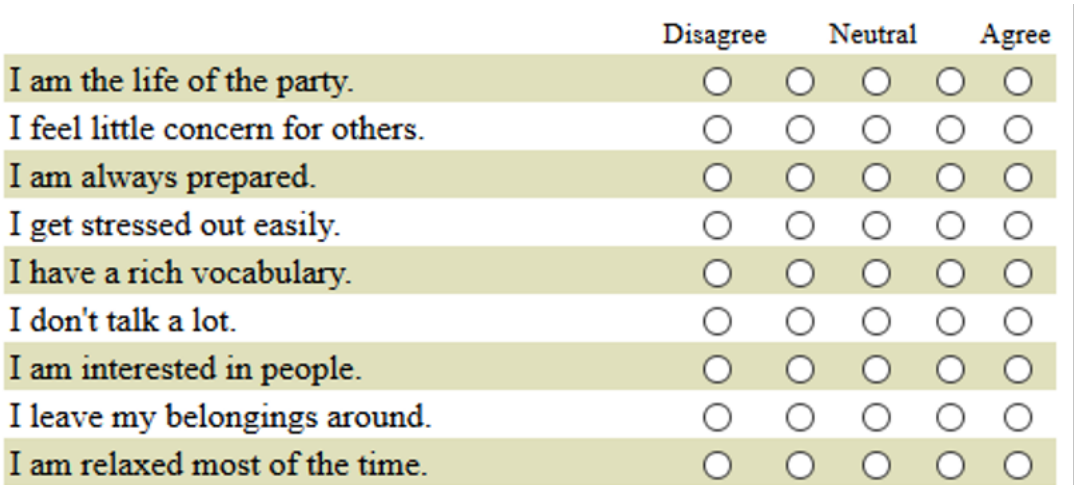
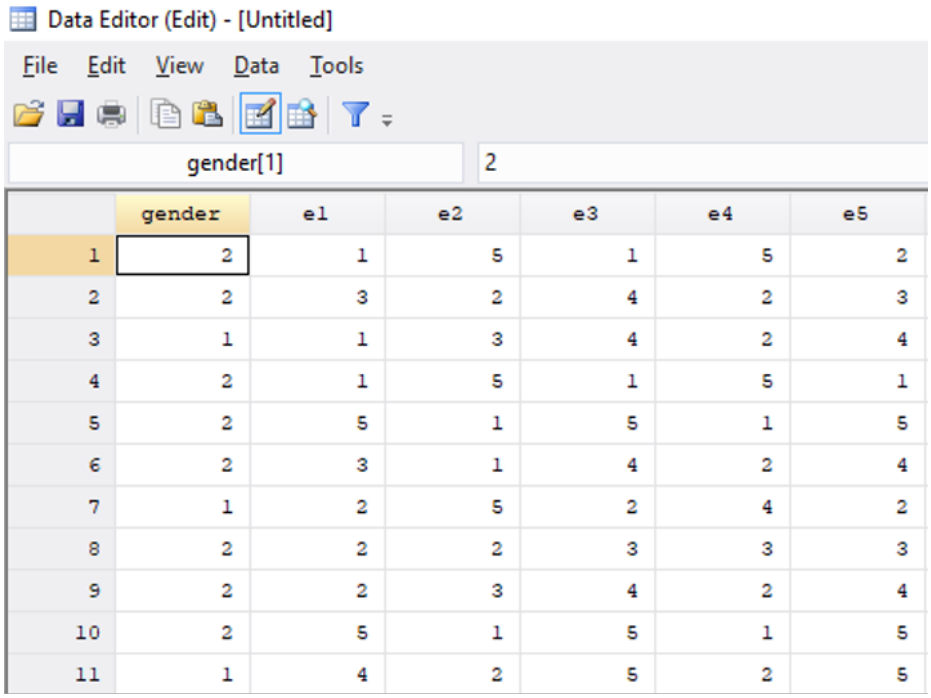


Figure 3.1

For now, I decide to focus on whether people are introverted or extroverted. Extroverts are outgoing and tend to enjoy interacting with others. Extroverts will tend to agree with the statement "I am the life of the party" while introverts will tend to agree with the item "I don't talk a lot."

I find a dataset that contains lots of responses to the Big Five personality questions as well as information on the gender of each respondent.[1] There are 10 different questions related to extroversion, and the dataset has one variable (column of data) for each of these 10 questions. The column labeled e1 shows responses to the item "I am the life of the party." A value of 1 means the respondent disagrees with this statement, while a 3 indicates neutral, and a 5 means they disagree.

[1] https://openpsychometrics.org/_rawdata/ (the file I used is called "BIG5.zip")



	gender	e1	e2	e3	e4	e5
1	2	1	5	1	5	2
2	2	3	2	4	2	3
3	1	1	3	4	2	4
4	2	1	5	1	5	1
5	2	5	1	5	1	5
6	2	3	1	4	2	4
7	1	2	5	2	4	2
8	2	2	2	3	3	3
9	2	2	3	4	2	4
10	2	5	1	5	1	5
11	1	4	2	5	2	5

Figure 3.2

For all of the odd-numbered extroversion questions (e1, e3, e5, etc.), agreement indicates extroversion. For the even-numbered items (e2, e4, e6, etc.), agreement indicates introversion. To create a single extroversion variable that combines responses from all 10 survey items, I create a tally, adding up all the values for odd-numbered questions and then subtracting the responses to the even-numbered questions. An extreme extrovert will have a 5 for all the odd-numbered questions and a 1 for all of the even-numbered ones, giving them a score of 20 (5x5-5x1=20). An extreme introvert will have a -20 since they will answer 1 to all the odd-numbered questions and 5 to all the even-numbered ones (5x1-5x5=-20).

Most people lie somewhere in the middle between introversion and extroversion:

Our gender variable was measured by asking respondents "What is your gender?" and they could choose from male, female, or other. In a moment, we'll consider those who responded "other," but for now, let's just look at those who chose either male or female.

3.1 Predicting extraversion using gender

If I want to describe differences in extraversion by gender in this dataset, I can compute the mean value of extraversion for males and for females. It turns out that males have an average extraversion of -0.46 while females' average level of extraversion is 0.53. Thus, the average

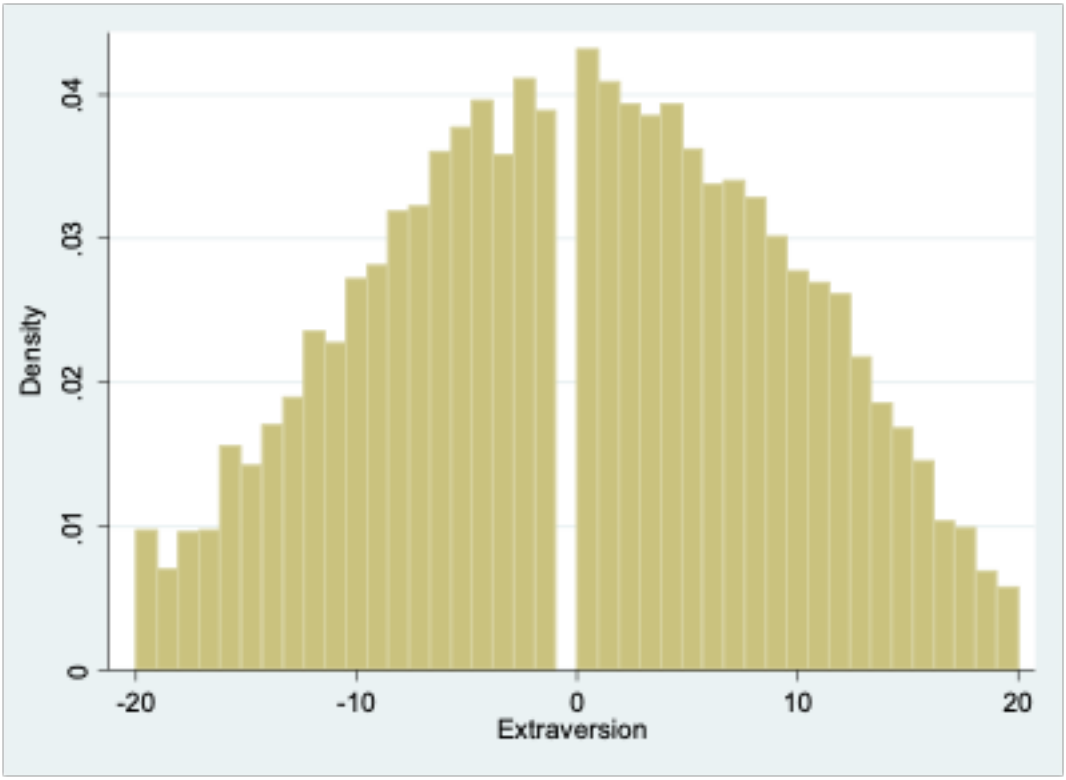


Figure 3.3

female is about 1-point more extraverted than the average male. But of course, there is lots of variation in extraversion among both groups:

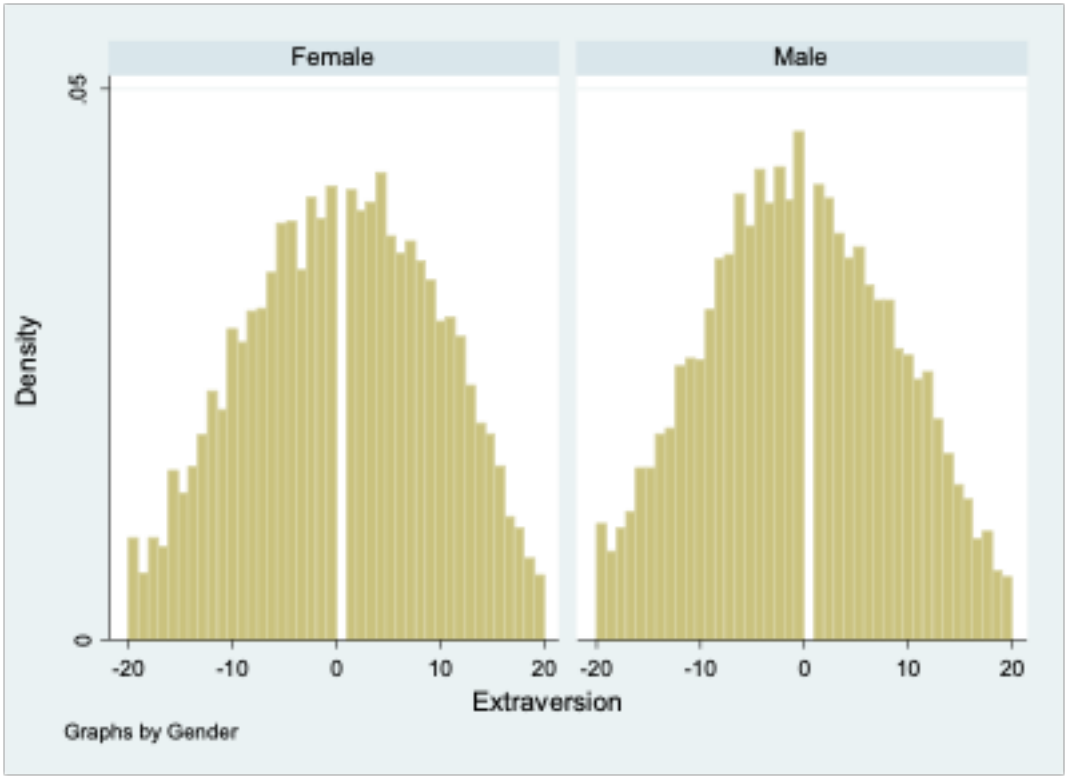


Figure 3.4

There are plenty of females who are introverts and plenty of males who are extroverts.

If you asked me to guess the extroversion level of someone and the only thing you told me about them was their gender, my best bet would probably be to guess the average extroversion level for someone of that gender. So for a female I knew nothing else about, I would guess their extroversion to be 0.53, while for a male I'd guess -0.46.

Social scientists use the **dependent variable** to describe the variable they're making a prediction about and **independent variable** to describe the variables that help them make that prediction. So in this example, extraversion is my dependent variable and gender is my independent variable.

When we're working with data, sometimes it's helpful to express how I would make a guess about a dependent variable (extraversion) based on other factors (gender) using a mathematical formula. In fact, this is exactly what we do when we run a regression. There are many ways I could write this formula, but I'll show just two for now. First, I could write:

$$\widehat{Extraversion} = 0.53 \times Female - 0.46 \times Male \tag{3.1}$$

Notice I've added a "hat" above the name of the variable *Extraversion*; this hat means that I'm making a guess about the value of that variable (I'm guessing the level of extraversion based on gender). The equation has two other variables *Female* and *Male*, and these two variables will take on a value of 1 if the person's gender is equal to the name of the variable and will otherwise take on a value of 0. For a female, *Female* will equal 1 and *Male* will equal 0, giving us:

$$\widehat{Extraversion} = 0.53 \times (1) - 0.46 \times (0) = 0.53$$

So our guess for the level of extroversion ($\widehat{Extraversion}$) of a female we know nothing about is 0.53.

For a male, our guess is:

$$\widehat{Extraversion} = 0.53 \times (0) - 0.46 \times (1) = -0.46$$

There's a second way I can write my formula, which will turn out to be more useful in the future when we come to consider multiple factors at the same time that might help us predict the value of a dependent variable. Rather than having two variables to represent gender in my equation, I can just use one:

$$\widehat{Extraversion} = 0.53 - 0.99 \times Male \tag{3.2}$$

In Equation 3.2, we start from female as our baseline. Notice that the first number we see (0.53) is our guess for the value of extraversion for a female. When we're considering a female, Male=0, so:

$$\widehat{Extraversion} = 0.53 - 0.99 \times 0 = 0.53$$

Thus, we get the right prediction for females from this equation, even though we didn't include a variable specifically for females. If we have a male, Male=1, so we get:

$$\widehat{Extraversion} = 0.53 - 0.99 \times 1 = -0.46$$

This is the same prediction we got before. Remember, I decided to initially just analyze respondents who selected either male or female. Since we are only considering two categories (male or female), and each respondent is either a male or a female, saying $Male = 1$ lets me know that $Female = 0$. It's actually repetitive in this context to both say that $Male = 1$

and *Female* = 0. Similarly, saying *Male* = 0 implies that *Female* = 1. So I can simplify my equation by just including one variable to indicate binary gender.

Notice that in Equation 3.2, the number next to *Male* is equal to the difference between the average level of extraversion for females and the average level for males (0.53-(-0.46)=0.99). This is because Equation 3.2 starts with females as the baseline, so to get our prediction for males, we have to adjust our baseline prediction by the average difference for males.

Equation 3.2 is also typically how we will arrange our equation when we're running a regression.

3.2 Prediction with more than two categories for gender

I now move beyond the gender binary and consider the "other" category in survey responses. I'll refer to this other category as "non-binary" gender. The average level of extraversion among those with non-binary gender is -5.66. So non-binary people tend to be quite a bit more introverted than those who identify as male or female. As with males and females, there is considerable variation among non-binary people:

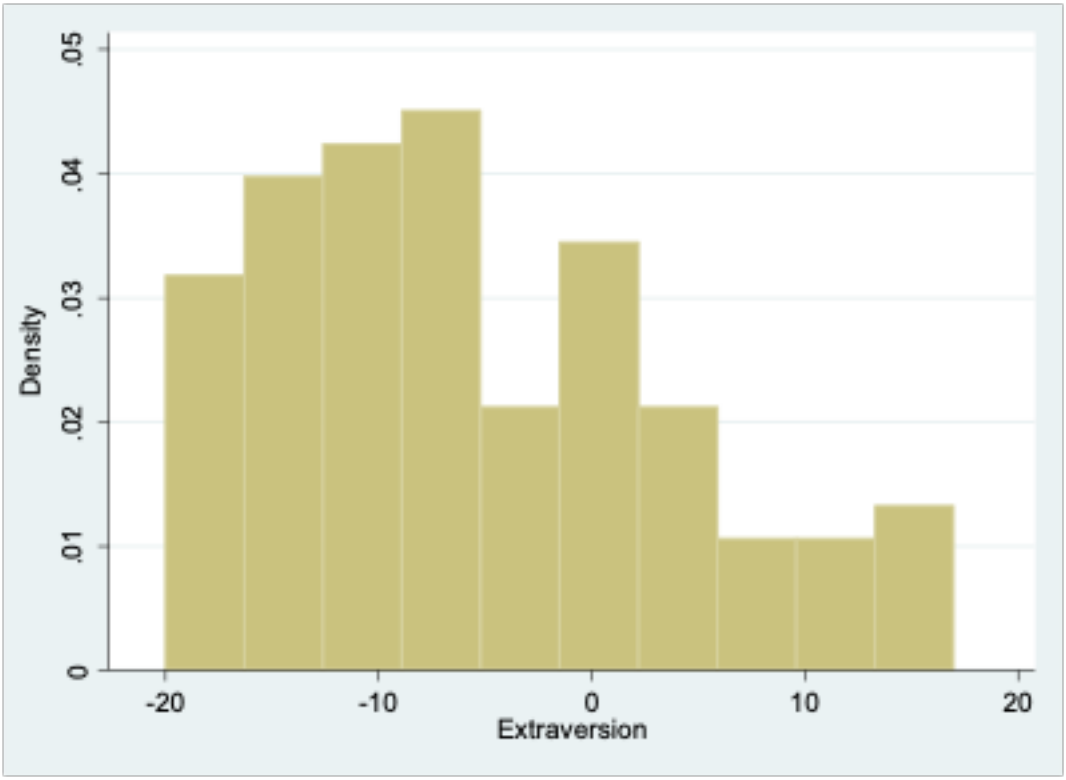


Figure 3.5

The number of non-binary respondents is relatively small (102), so it's not terrible surprising that this histogram looks a bit choppier than the ones we saw before.

Again, if we had to make a guess about the level of extraversion of someone, and all we knew about that person was that their gender was non-binary, we would probably want to guess the mean value among non-binary respondents (-5.66). Modifying Equation 3.1 to incorporate a third category is relatively straightforward:

$$\widehat{Extraversion} = 0.53 \times Female - 0.46 \times Male - 5.69 \times Other \tag{3.3}$$

For someone who identifies as female, we would plug in $Female = 1$, $Male = 0$, and $Other = 0$:

$$\widehat{Extraversion} = 0.53 \times (1) - 0.46 \times (0) - 5.66 \times (0) = 0.53$$

If someone identifies as non-binary, we would use $Female = 0$, $Male = 0$, and $Other = 1$:

$$\widehat{Extraversion} = 0.53 \times (0) - 0.46 \times (0) - 5.66 \times (1) = -5.66$$

We can also return to the format of Equation 3.2 but modify it to include the other category. This is how we will typically write our equation if we are doing a regression:

$$\widehat{Extraversion} = 0.53 - 0.99 \times Male - 6.19 \times Other \tag{3.4}$$

Now that there are three possible values for gender (female, male, and other), knowing the value of *Male* doesn't necessarily allow us to conclude what the vale of female is. If , the individual could identify as either female or non-binary. So we have to include a second variable. In this case, we chose to include the variable *Other*. If we know the values of *Male* and *Other*, we can always figure out the value of *Female* by process of elimination.

For a non-binary person, we plug in $Male = 0$, and $Other = 1$:

$$\widehat{Extraversion} = 0.53 - 0.99 \times (0) - 6.19 \times (1) = -5.66$$

When considering a female, we use $Male = 0$, and $Other = 0$:

$$\widehat{Extraversion} = 0.53 - 0.99 \times (0) - 6.19 \times (0) = 0.53$$

Equation 3.3 and Equation 3.4 communicate an equivalent method of making a prediction about extraversion based on gender; they just offer this information in two different formats.

Equation 3.4 might be a bit trickier to understand for now, but it will become very useful in the future.

Notice that we can talk about gender either as one qualitative variable with three possible values (female, male, or other), or we can talk about it as a series of three dummy variables (*Female*, *Male*, and *Other*) that can take each on a value of either 0 or 1. This can make things a bit confusing, but the important thing to remember is that when we have a qualitative variable with more than two categories, we'll need to break out the categories into a set of dummy variables for purposes of representing the qualitative variable in an equation.

However, as Equation 3.2 and Equation 3.4 illustrate, we don't necessarily need a dummy variable for every single category. Specifically, whenever we want to create an equation with a qualitative independent variable in a format like Equation 3.2 or Equation 3.4, the number of dummy variables should be equal to the number of categories minus one. Since our gender variable can take on three possible values in this example, we included two independent variables in Equation 3.4. No dummy variable is included for female, so we call female the **omitted category** or the **baseline category**. Remember, the first number in Equation 3.4 is 0.53, which represents our guess for females—the baseline category. If we instead had a qualitative variable with five categories, we would include four dummy variables in our equation.

4 Regression with Qualitative Dependent Variables

Suppose I want to build a model of voting. I decide to use the 2016 American National Election Studies[1] survey results to try to understand how race is associated with voting. Respondents in the 2016 survey were asked about who they voted for in 2012, and I'm going to focus on their 2012 voting patterns for now. Here are the distributions for my two main variables of interest:

```
```{stata}

. tab vote
PRE: RECALL OF LAST (2012) PRESIDENTAL  |
                            VOTE CHOICE |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                       1. Barack Obama |      1,728       56.58       56.58
                        2. Mitt Romney |      1,268       41.52       98.10
                      5. Other SPECIFY |         58        1.90      100.00
----------------------------------------+-----------------------------------
                                 Total |      3,054      100.00

. tab race
  PRE: SUMMARY - R SELF-IDENTIFIED RACE |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                1. White, non-Hispanic |      3,038       71.68       71.68
                2. Black, non-Hispanic |        398        9.39       81.08
3. Asian, native Hawaiian or other Paci |        148        3.49       84.57
4. Native American or Alaska Native, no |         27        0.64       85.21
                           5. Hispanic |        450       10.62       95.82
6. Other non-Hispanic incl multiple rac |        177        4.18      100.00
----------------------------------------+-----------------------------------
                                 Total |      4,238      100.00
```
```

[1] https://electionstudies.org/data-center/2016-time-series-study/

Notice that my dependent variable (vote) is qualitative. It can take on three possible values: voted for Obama, voted for Romney, or voted for other. I can build a simple set of regression models to see how race predicts vote choice. The key is to first convert each of the three categories for my dependent variable into its own dummy variable. I can accomplish this with the following code:

```
```{stata}
tab vote, gen(vote_)
```
```

I now have several new variables in my dataset that have names starting with "race_":

```
```{stata}
. tab vote_1

   vote==1. |
     Barack |
      Obama |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,326       43.42       43.42
          1 |      1,728       56.58      100.00
------------+-----------------------------------
      Total |      3,054      100.00

. tab vote_2

   vote==2. |
Mitt Romney |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,786       58.48       58.48
          1 |      1,268       41.52      100.00
------------+-----------------------------------
      Total |      3,054      100.00

. tab vote_3

   vote==5. |
      Other |
    SPECIFY |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,996       98.10       98.10
          1 |         58        1.90      100.00
```
```

```
------------+-----------------------------------
      Total |      3,054      100.00
```

I also convert my race variable into a set of dummy variables by running:

```{stata}
tab race, gen(race_)
```

I can then run three regressions, one for each value of my dependent variable. Let's start with voting for Obama (vote_1):

```{stata}
. reg vote_1 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =     76.29
       Model |  83.3981974         5  16.6796395   Prob > F        =    0.0000
    Residual |  662.426572     3,030  .218622631   R-squared       =    0.1118
-------------+----------------------------------   Adj R-squared   =    0.1104
       Total |  745.824769     3,035  .245741275   Root MSE        =    .46757

------------------------------------------------------------------------------
      vote_1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |   .4972868   .0281049    17.69   0.000     .4421802    .5523934
      race_3 |   .2078207   .0541766     3.84   0.000     .1015941    .3140472
      race_4 |   .1028423   .1353307     0.76   0.447     -.162507    .3681916
      race_5 |   .3135004    .032158     9.75   0.000     .2504466    .3765542
      race_6 |   .1042547   .0441427     2.36   0.018      .017702    .1908075
       _cons |    .480491   .0097901    49.08   0.000     .4612952    .4996868
```

Since our independent variable is qualitative, we have an omitted category. In this case, we've left category 1 (race_1) out of our regression, which indicates non-Hispanic white respondents. Our constant indicates that predicted value of the dependent variable when all independent variables are equal to zero. We can see this by writing out the regression equation:

$$\widehat{vote\_1} = .48 + .50race\_2 + .21race\_3 + .10race\_4 + .31race\_5 + .10race\_6 \qquad (4.1)$$

For non-Hispanic white respondents, race_1 equals one and all other race dummy variables equal zero, so we get:

$$\widehat{vote\_1} = .48 + .50(0) + .21(0) + .10(0) + .31(0) + .10(0) = .48$$

Remember, vote_1 is equal to zero if the respondent didn't vote for Obama, and it is equal to one if the respondent did vote for Obama. Our predicted value is neither zero nor one; instead, we get .48. This can be interpreted as indicating the probability of a one. In other words, a non-Hispanic white has a .48 probability of voting for Obama. We can also convert this probability to a percentage by moving the decimal place two spots to the right: a non-Hispanic white is estimated to have a 48% chance of voting for Obama, according to this model.

Now, let's look at the slope coefficients. The coefficient for black (race_2) equals .50. Thus, a one-unit increase in race_2 is associated with a .50-unit increase in vote_1. Let's break that down a bit to see if we can create a clearer interpretation. Since race_2 is a dummy variable and non-Hispanic white is the omitted category, a one-unit increase in race_2 correspondents to having a black respondent instead of a white respondent. And since our dependent variable is binary, we should think in terms of probabilities, which can be converted to percentages: a .50-unit increase in vote_1 means a 50 percentage-point increase in the probability of voting for Obama. So putting this altogether, we'd say: (non-Hispanic) black voters are 50 percentage points more likely to vote for Obama than (non-Hispanic) white voters, according to this model.

Similarly, Asian voters are 21 percentage points more likely to vote for Obama than (non-Hispanic) white voters. Native Americans are 10 percentage points more likely to vote for Obama than (non-Hispanic) white voters. Hispanics are 31 percentage points more likely to vote for Obama than non-Hispanic white voters. And voters identifying as multiracial or other race are 10 percentage points more likely to vote for Obama than (non-Hispanic) white voters. All of these differences are statistically significant, except for Native American versus white voters (probably because there are only 27 Native Americans in the sample, making the estimate of this difference very imprecise).

Let's move onto running a regression for the second category of our dependent variable:

```
```{stata}
. reg vote_2 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =     72.35
       Model |  78.6117037         5  15.7223407   Prob > F        =    0.0000
    Residual |  658.463395     3,030  .217314652   R-squared       =    0.1067
-------------+----------------------------------   Adj R-squared   =    0.1052
```
```

```
       Total |  737.075099     3,035  .242858352   Root MSE        =    .46617

------------------------------------------------------------------------------
      vote_2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |   -.483031   .0280207   -17.24   0.000    -.5379725   -.4280895
      race_3 |  -.2002027   .0540143    -3.71   0.000     -.306111   -.0942944
      race_4 |  -.0822373   .1349253    -0.61   0.542    -.3467917     .182317
      race_5 |  -.3014791   .0320617    -9.40   0.000     -.364344   -.2386142
      race_6 |  -.1344972   .0440105    -3.06   0.002    -.2207906   -.0482038
       _cons |    .498904   .0097607    51.11   0.000     .4797657    .5180423
```

Now we're looking at predictions of voting for Mitt Romney. Our constant is .50, indicating that a non-Hispanic white voter has a 50% chance of voting for Mitt Romney. The coefficient of -.48 for race_2 indicates that (non-Hispanic) black voters are 48 percentage points less likely to vote for Mitt Romney than (non-Hispanic) white voters. I won't go on to interpret the rest of the coefficients, but they follow the same pattern.

Finally, let's look at a regression with vote_3 as the dependent variable:

```
```{stata}
. reg vote_3 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =      2.23
       Model |   .20833556         5  .041667112   Prob > F        =    0.0490
    Residual |  56.6836275     3,030  .018707468   R-squared       =    0.0037
-------------+----------------------------------   Adj R-squared   =    0.0020
       Total |  56.8919631     3,035  .018745293   Root MSE        =    .13678

------------------------------------------------------------------------------
      vote_3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |  -.0142558   .0082213    -1.73   0.083    -.0303757    .0018642
      race_3 |   -.007618   .0158479    -0.48   0.631    -.0386917    .0234557
      race_4 |   -.020605   .0395873    -0.52   0.603    -.0982258    .0570158
      race_5 |  -.0120213    .009407    -1.28   0.201     -.030466    .0064234
      race_6 |   .0302425   .0129128     2.34   0.019     .0049238    .0555611
       _cons |    .020605   .0028638     7.19   0.000     .0149898    .0262202
```
```

This regression provides some insights into who supported third-party candidates in the 2012 election. First, our constant indicates that a non-Hispanic white voter has a 2% chance of voting third-party. (Non-Hispanic) black voters are one percentage point less likely to vote third-party than white voters, although this difference is only significant at the .10 level. The only other significant slope coefficient is for race_6, where we see that people who identify as multiracial or other race are estimated to be three percentage points more likely to vote third-party than (non-Hispanic) white respondents.

Now that we've run one regression for each category of our dependent variable, we've completed an analysis. Note that using regular linear regression (the reg function in Stata) is not the only way (or even necessarily the preferred way) to analyze a qualitative dependent variable. There are other models (e.g., multinomial logistic regression) that are specifically designed to be used with a qualitative dependent variable. However, using simple linear regression is a good way to get started looking at qualitative variables if you haven't learned these fancier models and how to properly interpret them.

One final thing I want to show you is that our results will be in a slightly different format but will be in one sense equivalent if we decide to use a different category as our omitted category when using a qualitative independent variable. Let's say we want to make black (race_2) our reference category. Compare the following results to what we saw near the top of this page:

```{stata}
. reg vote_3 race_1 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =      2.23
       Model |   .20833556         5  .041667112   Prob > F        =    0.0490
    Residual |  56.6836275     3,030  .018707468   R-squared       =    0.0037
-------------+----------------------------------   Adj R-squared   =    0.0020
       Total |  56.8919631     3,035  .018745293   Root MSE        =    .13678

------------------------------------------------------------------------------
      vote_3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_1 |   .0142558   .0082213     1.73   0.083    -.0018642    .0303757
      race_3 |   .0066378    .017388     0.38   0.703    -.0274557    .0407313
      race_4 |  -.0063492   .0402287    -0.16   0.875    -.0852274     .072529
      race_5 |   .0022345   .0118186     0.19   0.850    -.0209387    .0254077
      race_6 |   .0444983   .0147623     3.01   0.003      .015553    .0734435
       _cons |   .0063492   .0077064     0.82   0.410    -.0087611    .0214595
```

Now, our constant tells us that a black voter has a .6% chance of voting third-party. This is the same prediction we would get from our prior model where race_1 was the omitted category: to

find our prediction for black voters from the prior results we would have added the coefficient for race_2 (-.014) to the constant (.021), yielding .6% or .006 (or .007 if we use the rounded numbers shown in parentheses).

The coefficient for race_1 tells us about how white voters differ from black voters. Notice that the p-value is exactly the same as what we saw in the prior table for race_2, and the coefficient for race_1 in this table is the same as the coefficient for race_2 in the prior table, except the sign has changed. That's because comparing black to white is the same as comparing white to black, except that we're going in the opposite direction.

You can go on to play around with these two sets of results more on your own if you'd like. Both regression equations will yield the same prediction for a voter of any given race. The difference lies only in the starting point, as represented fby the constant. However, the p-values will usually differ because they are describing a different comparison (e.g., comparing Asian to black in this table versus comparing Asian to white in the prior table). Thus, it doesn't really matter which category you pick as your omitted category, except that you may care more about some comparisons than others. You can also run the same regression multiple times but with different omitted categories so that you can get the p-values for a full set of comparisons across groups.