Datasets Codebook

Sociology 312, University of Oregon
Prof. Gullickson

Add Health

This data comes from the National Longitudinal Study of Adolescent to Adult Health (Add Health), conducted by the Carolina Population Center at UNC-Chapel Hill and supported by a grant from the National Institute of Child Health and Human Development. The first wave of the study which we are using surveyed adolescents between 7th and 12th grade in school in the 1994-95 school year.

One of the particularly valuable features of the Add Health survey is that many respondents were in the "saturation sample" which sampled *all* students at 16 schools. In this saturation sample, students were asked about who were their friends and sexual partners, which allows researchers to construct network maps of adolescent social systems.

We will use this saturation sample to look at a various basic measure of that network that estimates students' popularity. This measure, which is called "in degree" in the network analysis literature, measures the number of times a student was nominated as a friend by other students in the school. We will treat it as a simple proxy measure of a student's popularity. We can then look at what other student characteristics were positively or negatively associated with a student's popularity.

Here is a full description of all variables in the dataset that we will use.

- **indegree**: The number of friend nominations received by other students at the same school. This is the measure of popularity that we will use.
- **race**: A six-category nominal variable indicating the race that the student best thought described them when asked to choose a single race: white, black, Latino, Asian, American Indian, other.
- **sex**: Add Health reports this as a student's "biological" sex. Students were only reported as male or female.
- **grade**: current grade of the student as a quantitative variable.
- **psuedoGPA**: Students were asked for the most recent letter grade in four course types: math, language arts, science, and math. This variable was constructed by calculating GPA from those responses.
- **honorsociety**: A true/false variable for whether a student was in honor society or not.
- **alcoholuse**: A true/false variable that is true if the student reported drinking at least once or twice a month in the last twelve months.
- **smoker**: A true/false variable that is true if student smoked more than 5 cigarettes in the past 30 days.
- **bandchoir**: a true/false variable that is true if the student was in band or choir.
- **academicclub**: a true/false variable that was true if the student was in an academically-oriented club such as math club, book club, etc.
- **nsports**: The number of different school sports a student reported participating in. Students who reported more than six sports were top-coded at the value of six.
- **parentinc**: Parent's household income measured in $1000's of dollars.

Crimes

The crimes data contain information on crime rates and demographic variables for all fifty US states and the District of Columbia. The crime rates are for the year 2010 and come from the FBI's Uniform Crime Reports (UCR). The UCR is a program where local law enforcement agencies all report crime statistics to the FBI and these are aggregated into final crime statistics. For our purposes, we are dividing crimes into two main categories of violent and property crime.

The demographic characteristics come from the American Community Survey (ACS) between the years 2008 and 2012. The ACS is an annual sample of the US population. To get a large enough sample in each state to calculate correct statistics (with little sampling error), I combine five years of data that are "centered"" on 2010.

- **Violent**: violent crimes per 100,000 population within each state. This includes the crimes of murder, rape, robbery, and aggravated assault. By dividing the number of crimes by the population size, we avoid the problem of larger population states having more crimes because of a larger population. This is often called the crime "rate."
- **Property**: property crimes per 100,000 population. This includes the crimes of burglarly, larceny, and motor vehicle theft.
- **MedianAge**: Median age of a state's population.
- **PctMale**: Percent of a state population that is male.
- **PctLessHS**: The percent of the state population over the age of 25 without a high school diploma.
- **MedianIncomeHH**: Median household income in a state. This is measured in thousands of dollars (i.e. 35 means $35,000). We are taking the income of each household (meaning all members of that household combined) rather than individual level income. For most purposes, this is thought to be a better measure because consumption and savings are typically organized at the household level.
- **Unemployment**: Unemployment rate in the state. The unemployment "rate" is really just a percentage. Its the percentage of individuals who are not working but want to work among all those in the labor force (those who are working or looking for work).
- **Poverty**: Poverty rate in the state. The poverty "rate" is also really just a percentage. It is the percent of individuals living below the poverty line. The poverty line is a number developed by the federal government. It was originally developed in the 1960s and is adjusted for inflation every year. Many people critique the poverty line as being too low because it has not kept pace with increases in the consumer price index.
- **Gini**: A measure of income inquality in the state. The gini coefficient is a widely used measure of how unequally income is distributed. If gini is zero, then everyone has exactly the same income. If gini is 100, then one person makes all the money and everyone else zero. The higher the gini coefficient, the more income inequality exists. You can get a more detailed description here.

Movies

The movie data contain information about 2,612 movies produced between 2001 and 2013. The data come from the Open Movie Database, which itself contains data from the Internet Movie Database and Rotten Tomatoes. To simplify our analyses, I have limited the analysis to movies that played in the US and received 10 or more reviews. I have also excluded all shorts, documentaries, foreign language films, and movies that received an NC-17 rating or were unrated. Here are the variables we have for each movie:

- **Year**: The calendar year of the film's release.
- **Rating**: The movie's maturity rating (G, PG, PG-13, R).
- **Runtime**: The length of the movie in minutes.
- **Oscars**: The number of Oscar awards that the movie received. This includes Oscars that go to individiual actors (leading and supporting), as well as more general awards (best screenplay, editing, cinematography, etc.), and best picture overall.
- **TomatoMeter**: The tomato meter is the percent of reviews that are judged to be positive by Rotten Tomatoes staff. This metric goes from 0 to 100 percent. Note that movies are spread out pretty evenly across the range of this variable, something we call a "uniform" distribution. Note, that this method makes no distinction between how positive a positive review was or how negative a negative review was, so its perfectly possible for two movies with the same Tomato Meter to be viewed very differently by reviewers.
- **TomatoRating**: The tomato rating is a combination of all reviews where the review used some kind of numeric rating (e.g. 3 out of 4 stars, 7 out of 10). Rotten Tomatoes "normalizes" these scores so that they are all recorded on the same basis. The scale of this normalized score goes from 1 to 10. Unlike

the Tomato Meter, this scale should be capable of distinguishing how strongly positive or negative the review was.
- **BoxOffice**: The box office returns for the movie in millions of US dollars.
- **Genre**: The genre of the film. This is a tricky variable to create. In actuality, movies could be listed as multiple genres in the original dataset, with twenty different genres to choose from. For example, "No Country for Old Men" is listed in the genres of crime, drama, and thriller while "Lord of the Rings: Return of the King" is listed as action, adventure, and fantasy. This is probably the best way to treat genres, but for our purposes it adds a lot of complexity. Therefore, I have recoded movies into a single "best" genre based on a decision rule where certain genres trump all others on an ordered basis. For example, comedy trumps romance, so romantic comedies will always show up in this dataset as comedies. The ordering of this system is Animation > Family > Musical > Horror > SciFi/Fantasy > Comedy > Romance > Action > Thriller > Mystery > Drama > All Others. For the most part, this system works well, but you may notice some odd disrepancies for a few movies.

Politics

This data comes from the 2016 American National Election Study (ANES). The ANES is a survey of the American electorate that is conducted every two years. The study collects information on a variety of political attitudes and voting behaviors. For our purposes, we are going to primarily look at respondent's vote for president and attitudes on three issues: (1) birthright citizenship, (2) gay marriage, and (3) global warming. The variables we will look at are:

- **brcitizen**: Respondents were asked whether they would support a proposal to change the US Constitution to remove birthright citizenship (citizenship automatically granted to individuals born in the US regardless of their parent's citizenship status). Respondents could either favor, oppose, or neither favor or oppose.
- **gaymarriage**: Respondents were asked for their position on gay marriage and were given the choices of "no legal recognition", "civil union (but no marriage)", "support gay marriage."
- **globalwarm**: A question on whether the respondent believes that anthropogenic global warming is happening. I constructed this variable from two separate questions. The first question asks whether respondents think that global warming has been happening with the options being that it "probably has" or "probably has not." The second question asks whether respondents thought that global warming was caused by human activity (either entirely or partially). I combine these into a single dichotomous variable where individuals either think the earth is warming from human activity or that it is not warming from human activity, where the latter category includes people who think it isn't warming at all and people who think it is warming but not because of human activity.
- **party**: The political party with which the respondent identifies. This does not necessarily mean that a respondent is officially registered with a given party.
- **relig**: The respondent's religion. This category is based on the combination of people's statement about the kind of services they typically attend along with several non-exclusive yes/no questions about their religion (e.g. evangelical, pentecostal, agnostic, aethist).
- **age**: The age of the respondent.
- **gender**: The respondent's self-reported gender, recorded as "Male","Female", or "Other."
- **race**: the racial identification of the respondent. Respondents could write in multiple races, but to keep it simple, we will combine the small number of individuals who reported multiple races with those who listed "Other" as their race.
- **educ**: The education of the respondent. This is recorded as an ordinal variable. The "Some college" response indicates individuals who have attended college (including 2-year programs) but have not earned a BA.
- **income**: The family income of the respondent in 1000s of dollars. Respondents did not give actual dollar amounts here but rather indicated which bracket of income (e.g. $20,000-30,000) they fell within. For the purposes of our class, I randomly select an actual value within this bracket for each respondent.
- **workstatus**: The work status of the respondent. Respondents could either be working, unemployed, or out of the labor force. The last category refers to people who are not employed and not currently

looking for work, whereas unemployed indicates a person who is not employed an is currently looking for work.

- **military**: Whether the respondent has ever served or is currently serving in the US military.

Sex

The sex data come from a special supplemental questionnaire that was added to the General Social Survey (Links to an external site.)Links to an external site. (GSS) in 2004. The GSS is a survey of attitudes that is conducted every two years by the National Opinion Research Council (NORC). In the 2004 supplement, respondents were asked questions about their sexual behavior. We will be looking specifically at respondents reported frequency of sexual activity and its relationship to demographic characteristics such as age, education, and marital status. Here are the variables we will look at:

- **sexf**: A quantitative variable indicating the frequency of sexual activity as the number of sexual encounters per year. The sequal frequency response was coded as an ordinal scale variable in which respondents were given a set of options from less to more sexual activity in the previous year. For our purposes, I have recoded the ordinal sexfreq variable into a quantitative variable by giving everyone the midpoint number of sexual acts per year based upon their answer. For example, individuals who said (2 or 3 times a month) were given a value of $2.5 * 12 = 30$. This will allow us to use sexual frequency as a dependent variable in regression models.
- **age**: The age of the respondent. The GSS only surveys adults aged 18 years and older.
- **gender**: The gender of the respondent.
- **educ**: Years of education for the respondent.
- **marital**: Marital status of the respondent: Never married, married, divorced, widowed, separated.
- **relig**: Religious affiliation of the respondent. Protestants have been divided into "mainline" and "fundamentalist" based on a coding of specific denominations used by the GSS.

Titanic

The titanic data contain information on all 1,309 passengers aboard the Titanic. The data do not include information about the crew. The data primarily come from the online database, Encyclopedia Titanica (Links to an external site.)Links to an external site.. Here are the variables we will look at:

- **survival**: Did the passenger survive?
- **sex**: The reported sex of the passenger.
- **age**: The age of the passenger. This variable is reported in whole numbers for those over one year old and as a decimal (based on months of age) for infants under a year of age.
- **agegroup**: A categorical variable indicating whether the person was an adult or a child. I have constructed this variable from the age variable. The cutoff for adults is sixteen years of age.
- **pclass**: There were three passenger classes: First, second, and third (also known as steerage). To give some pop culture references, Rose was first class, and Jack was third class. Most of the passengers were in third class.
- **fare**: The fare paid for the ticket, measured in British pounds.
- **family**: The number of family members traveling with the passenger. These family members can either be parents, spouses, siblings, or children.