STATS 205: Midterm Project Proposal

Brian Liu

5/3/2019

1. A simple and clear exposition of the problem I am addressing

Suicide is prevalent in many times and places around the world, but many places and times have different suicide rates. When it comes to suicide, there are many potential factors or attributes that may be correlated with an increased risk of suicide, such as:

- a person's sex
- the age group a person belongs to
- the generation a person was born in

The goal is to find significant correlations between these factors and suicide rates: that is, does x factor positively predict suicide rate?

We will use the statistical techniques of nonparametric bootstrap and parametric bootstrap methods to aid in prediction, perhaps with linear regression as well, and use cross-validation to test if, given new data for a population, this population is at risk of suicide. In other words, predict if the suicide rate would be abnormally or significantly high, and then compare the performance between the two methods (nonparametric and parametric).

The simple inspiration is suicide prevention: If we can identify the factors that correlate positively with, or predict high suicide rates, then we can target our suicide prevention efforts towards populations with those high-risk factors or attributes.

2. Description of the data

The dataset we have is a `.csv` file containing data from four other datasets:

- United Nations Development Program. (2018). Human Development Index (HDI)
- World Bank. (2018). World Development indicators: GDP (current US$) by country:1985 to 2016.
- [Szamil]. (2017). Suicide in the Twenty-First Century [dataset].
- World Health Organization. (2018). Suicide prevention.

The data contains columns with factors/indicators/attributes including:

- `country`
- `year`
- `sex`
- `age` (e.g. 15-24 years, 35-54 years, 75+ years, etc.)
- `suicides_no` (number of suicides)
- `population` (of people in country)
- `country-year` (e.g. `Albania1992`)
- `HDI for year` (Human Development Index)
- `gdp_for_year`

3. Review of available methods for such data.

Overview of resampling

In statistics, resampling is any variety of methods for doing one of the following:

1. Estimating the precision of sample statistics (medians, variances, percentiles by using subsets of available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping)

– Resampling (statistics)

Overview of plug-in principle

In statistics, the plug-in principle is the method of estimation of functionals of a population distribution by evaluating the same functionals at the empirical distribution based on a sample. For example, when estimating the population mean, this method uses the sample mean; to estimate the population median, it uses the sample median; to estimate the population regression line, it uses the sample regression line. It is called a principle because it is too simple to be otherwise, it is just a guideline, not a theorem.

– Plug-in principle

Overview of bootstrap

In statistics, bootstrapping is any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods... Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution. One standard choice for an approximating distribution is the empirical distribution function of the observed data. In the case where a set of observations can be assumed to be from an independent and identically distributed popuation, this can be implemented by constructing a number of resamples with replacement, of the observed dataset (and of equal size to the observed dataset)... It may also be used for constructing hypothesis tests. It is often used as an alternative to statistical inference based on the assumption of a parametric model when that assumption is in doubt, or where parametric interference is impossible or requires complicated formulas for the calculation of standard errors.

– Bootstrapping (statistics)

Parametric bootstrap

Using the parametric bootstrap statistical method means making an assumption about the underlying distribution of the population. For example, we might assume that the underlying distribution of the population is a normal distribution.

Nonparametric bootstrap

In nonparametric bootstrap, we make no assumptions about the distribution of the underlying population.

Cross-validation

Essentially, we use some of our data (training set) to train a bootstrap predictor(s), then use some other data (test set) as a test of accuracy of our predictor(s).

4. Advantages and disadvantages of those methods, in particular, parametric versus nonparametric methods

Parametric bootstrap

- **Both advantage and disadvantage**: We would make an assumption about the model / population.
- **How it could be an advantage**: If we correctly pick a model that accurately describes the population, it could be more accurate in prediction.
- **How it could be a disadvantage**: If we don't pick the right model, it could be much less accurate.

Nonparametric bootstrap

- **Advantage**: No assumptions about the model or population is made. This is more accurate if we don't already know the category of the underlying distribution that the population falls under.
- **Disadvantage**: We lose the fine-tunedness that parametric bootstrap would give us if we *did* know what category of distribution the population falls under).

Cross validation

- **Advantage**: We can test the accuracy of our predictors using data we already have; that is, we don't need to collect more data to assess our predictors.
- **Disadvantage**: This assumes we don't have the ability to collect more data.

5. A proposed solution to the problem using the nonparametric method

The solution to the problem would consist of following a few steps:

1. For each column of data, produce confidence intervals using both statistical methods (parametric and nonparametric bootstrap), training on some of the data (cross-validation).
2. Find significant correlation for each column / category of data vs. suicide rate.
3. Cross validate on test data.
4. Repeat many times using different subsets for testing and training (cross-validation).
5. Compare results.

6. A flowchart of the various tasks to complete the project

Data collection

I already have a dataset, but collecting new data would probably consist of:

- Collecting data on suicides in a particular year(s), country(s), or demographic(s)
- Compiling the data into a `.csv.` file

This is probably how the four sources listed above collected their data.

Programming

- Use R package `bootstrap` to produce parametric and nonparametric bootstrap confidence intervals
- Possibly use `pandas` library in Python for cross-validation of bootstrap intervals

Simulation

Cross-validation *is* a simulation or test in some sense.

Potential resources for future work

- Suicide rates overview 1985 to 2016. Compares socio-economic info with suicide rates by year and country
- Cross-validation and the Bootstrap

References

- Suicide Rates Overview 1985 to 2016 | Kaggle. Compares socio-economic info with suicide rates by year and country
- United Nations Development Program. (2018). Human Development Index (HDI)
- World Bank. (2018). World Development indicators: GDP (current US$) by country:1985 to 2016.
- [Szamil]. (2017). Suicide in the Twenty-First Century [dataset].
- World Health Organization. (2018). Suicide prevention.
- Bootstrapping (statistics)
- Resampling (statistics)