STATS 205: Final Project Write-Up

Brian Liu

6/14/2019

1. Background of the data and why it is interesting or important

The data we are using is the data from WHO suicide statistics from Kaggle. This gives population-based statistics on suicide rate (Szamil 2018).

The reason this data is interesting and important is that 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?

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. Explanation of the method studied and its properties

We will use the statistical techniques of nonparametric bootstrap and parametric bootstrap methods to aid in prediction, with linear regression as well (Kendall coefficient), 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).

Bootstrapping

> 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 (Efron and Tibshirani 1993; Efron 2003). This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods. Generally, it falls in the broader class of resampling methods("Bootstrap Methods," n.d.).
>
> 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 population, 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 inference is impossible or requires complicated formulas for the calculation of standard errors.

Nonparametric vs. Parametric bootstrap

Whereas nonparametric bootstraps make no assumptions about how your observations are distributed, and resample your original sample, parametric bootstraps resample a known distribution function, whose parameters are estimated from your sample. These bootstrap estimates are either used to attach confidence limits nonparametrically - or a second parametric model is fitted using parameters estimated from the distribution of the bootstrap estimates, from which confidence limits are obtained analytically. The advantages and disadvantages of this approach, compared to nonparametric bootstrapping, can be summarised as follows.

In the nonparametric bootstrap, samples are drawn from a discrete set of n observations. This can be a serious disadvantage in small sample sizes because spurious fine structure in the original sample, but absent from the population sampled, may be faithfully reproduced in the simulated data. Another concern is that because small samples have only a few values, covering a restricted range, nonparametric bootstrap samples underestimate the amount of variation in the population you originally sampled. As a result, statisticians generally see samples of 10 or less as too small for reliable nonparametric bootstrapping.

Small samples convey little reliable information about the higher moments of their population distribution function - in which case, a relatively simple function may be adequate.

Although parametric bootstrapping provides more power than the nonparametric bootstrap, it does so on the basis of an inherently arbitrary choice of model. Whilst the cumulative distribution of even quite small samples deviate little from that of their population, it can be far from easy to select the most appropriate mathematical function a priori. Maximum likelihood estimators are commonly used for parametric bootstrapping despite the fact that this criterion is nearly always based upon their large sample behaviour.

Choosing an appropriate parametric error structure for a statistic based upon small samples can be awkward to justify. Bootstrap t statistics present an additional problem, partly because of problems in estimating standard errors analytically, partly because of difficulties in working out a suitable number of degrees of freedom for your pivot's (presumed, but often large-sample-based) distribution.

So although parametric bootstrapping can be relatively straightforward to perform, and may be used to construct confidence intervals for the sample median of small samples, the bootstrap and estimator distribution functions are often very different. In addition, confidence limits may enclose invalid parameter values, and the coverage error is no better than nonparametric intervals.

Confusingly, whilst the parametric bootstrap is sometimes described as a basic bootstrap, resampling residuals is sometimes referred to as being 'semi parametric' - which is also used to describe test-inversion and smoothed sample bootstraps. Resampling residuals is most popularly used to obtain bootstrap confidence intervals for regression coefficients, for example in nonparametric regression. ("A Parametric or Non-Parametric Bootstrap?" n.d.)

Linear regression - Kendall rank correlation coefficient

In statistics, the Kendall rank correlation coefficient, commonly referred to as Kendall's tau coefficient (after the Greek letter τ), is a statistic used to measure the ordinal association between two measured quantities. A tau test is a non-parametric hypothesis test for statistical dependence based on the tau coefficient.

It is a measure of rank correlation: the similarity of the orderings of the data when ranked by each of the quantities. It is named after Maurice Kendall, who developed it in 1938 (Kendall 1938), though Gustav Fechner had proposed a similar measure in the context of time series in 1897 ("Measures of Association for Ordinal Data," n.d.).

Intuitively, the Kendall correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully different for a correlation of -1) rank between the two variables.

Both Kendall's τ and Spearman's ρ can be formulated as special cases of a more general correlation coefficient.

Cross validation

Cross-validation, sometimes called rotation estimation (Geisser 1993) ("A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," n.d.) (Devijver and Kittler 1982), or out-of-sample testing is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (called the validation dataset or testing set). (2 et al., n.d.) ("Newbie Question: Confused About Train, Validation and Test Data!" n.d.). The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias (Cawley and Talbot, n.d.) and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, in most methods multiple rounds of cross-validation are performed using different partitions, and the validation results are combined (e.g. averaged) over the rounds to give an estimate of the model's predictive performance.

In summary, cross-validation combines (averages) measures of fitness in prediction to derive a more accurate estimate of model prediction performance.(Seni and Elder 2010)

3. Data analysis or simulation study

We will use the crude rate of suicide per 100,000 people.

This analysis provides information on age-standardized rates...

Filter and save countries with missing suicide rate.

After filtering countries with missing suicide rate, take a random sample of 100 countries and make sure each continent has approximately equal countries.

Filter countries by continent:

Let us find out which continents are counted:

Therefore,

$$\frac{100 \text{ countries}}{6 \text{ continents}} \approx 16 \text{ to } 17 \text{ countries per continent}$$

we should randomly sample 17 countries from each continent.

Notably, there are countries that are not on any of the listed continents. Let us see which ones those are:

Let us make the choice not to include these countries in the analysis, since there are only two countries.

We create six dataframes, filtered by list of countries for each continent.

This text links to very important information about why a `for` loop doesn't print anything.[1]

Link to Pandoc Markdown formatting

Randomly sample 17 countries from each continent:

Since there are only 5 countries in Oceania and 12 countries in Africa, we will use all 5 countries of Oceania and all 12 countries of Africa.

Let's see the countries that we will be sampling:

Let's filter the original dataframe only to include countries that we have sampled:

4. Interpretation of the results or discussion

5. References

2, Alexander GalkinAlexander Galkin 5, mohsen najafzadehmohsen najafzadeh 2, innovIsmailinnovIsmail 75753, Ryan ZottiRyan Zotti 3, Frank HarrellFrank Harrell 56.9k4115247, Yu ZhouYu Zhou 22122, et al. n.d. "What Is the Difference Between Test Set and Validation Set?" *Cross Validated*. https://stats.stackexchange.com/questions/19048/what-is-the-difference-between-test-set-and-validation-set/19051#19051.

"A Parametric or Non-Parametric Bootstrap?" n.d. *Parametric or Non-Parametric Bootstrap*. https://influentialpoints.com/Training/nonparametric-or-parametric_bootstrap.htm.

"A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." n.d. *ACM Digital Library*. Morgan Kaufmann Publishers Inc. https://dl.acm.org/citation.cfm?id=1643047.

"Bootstrap Methods." n.d. *From Wolfram MathWorld*. http://mathworld.wolfram.com/BootstrapMethods.html.

Cawley, Gavin C., and Nicola L. Talbot. n.d. "On over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation." *Journal of Machine Learning Research*. http://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf.

Devijver, Pierre A., and Josef Kittler. 1982. *Pattern Recognition: A Statistical Approach*. Sung Kang.

Efron, Bradley. 2003. *Second Thoughts on the Bootstrap*. Department of Biostatistics, Stanford University.

Efron, Bradley, and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman; Hall.

Geisser, Seymour. 1993. *Predictive Inference: An Introduction*. Chapman & Hall.

Kendall, M. G. 1938. "A New Measure of Rank Correlation." *Biometrika* 30 (1/2): 81. https://doi.org/10.2307/2332226.

[1] Basically, `for` loops are functions themselves. R prints out the result of a command automatically, but functions are not inherently a command, and since `for` loops are functions, nothing will be printed. The solution is to have `print(command())` within the `for` loop to get output for your `for` loop. You will never again spend hours trying to find out why a `for` loop doesn't print anything because you're no longer an R newbie.

"Measures of Association for Ordinal Data." n.d. *Measures of Association*, 64–85. https://doi.org/10.4135/9781412984942.n5.

"Newbie Question: Confused About Train, Validation and Test Data!" n.d. *Newbie Question: Confused About Train, Validation and Test Data! | Heaton Research*. https://web.archive.org/web/20150314221014/http://www.heatonresearch.com/node/1823.

Seni, Giovanni, and John F. Elder. 2010. "Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions." *Synthesis Lectures on Data Mining and Knowledge Discovery* 2 (1): 1–126. https://doi.org/10.2200/s00240ed1v01y200912dmk002.

Szamil. 2018. "WHO Suicide Statistics." *Kaggle*. https://www.kaggle.com/szamil/who-suicide-statistics.