Permutation Hypothesis Testing and Bootstrapping in Regression Model

Anna Shchiptsova

September 30, 2016

International Institute for Applied Systems Analysis, Schlossplatz 1, A-2361 Laxenburg, Austria

This document outlines the algorithm of permutation hypothesis testing and the bootstrap algorithm of parameter estimation in regression analysis.

CONTENTS

1 Software

No installation needed. All packages are standalone java applications.
Requires JRE 1.8 installed on the target machine.

URL for download:
http://www.iiasa.ac.at/web/home/research/researchPrograms/AdvancedSystemsAnalysis/land-use-spatial-analysis.html

Name	regression-tests-1.0.0-standalone.jar
Type	jar package
Summary	Library for spatial statistical analysis with resampling
Version	1.0.0
License	MIT, http://opensource.org/licenses/MIT
Imports	Clojure 1.8.0, https://clojure.org/; Incanter 1.5.7, http://incanter.org/
Command line options	-t, --trace Print stack trace
	-h, --help Print command help
Author and maintainer	Anna Shchiptsova, shchipts@iiasa.ac.at

2 Model

Let us consider a geographic region consisting of n administrative units. Suppose that we have panel data (X, y) collected in every administrative unit of the region. Here, X is a $n \times (p+1)$ matrix of the explanatory variables and y is a $n \times 1$ observable vector of the response. Each column X^i consists of the sample observations on a single explanatory variable.

In general, we want to relate the response variable to available explanatory factors in an administrative unit based on the reported panel data. For this purpose, we put forward a multiple regression model in the following form

$$\begin{gathered} y = X\beta + \varepsilon \\ \varepsilon_1, \dots, \varepsilon_n \sim F(0, \sigma^2) \end{gathered} \tag{1}$$

where $\beta = \left(\beta_0, \beta_1 \dots, \beta_p\right)^T$ is a $(p+1) \times 1$ vector of the unknown model parameters to be estimated from the data using the ordinary least squares method. By assumption, X^1 is identically 1, so that the regression equation has an intercept β_0. The error term ε is a $n \times 1$ vector of independent identically distributed errors with common distribution F having mean 0 and finite variance σ^2. Both F and σ^2 are unknown.

3 Permutation hypothesis testing

3.1 Theoretical background

A permutation test (Fisher, 1935) computes the probability for a null hypothesis being tested, whether the original test statistic of interest is a typical element of the set of statistics derived from the given observations by an appropriate class of data reordering. The basic idea of the test is free of the assumption on the exact distribution of the test statistic; rather the reference distribution is generated from the drawn permutation sample.

We run a permutation test with the chosen test statistic R^2 to assess the overall significance of model (1). The null hypothesis of the test states that the response variable has no linear relationship with the given explanatory variables, that is, $H_0: \beta_i = 0$ for all $i = 1 \dots p$ and the alternative hypothesis in the case of the two-tailed test is $H_1: \exists i\ \beta_i \neq 0\ (i = 1 \dots p)$. If the null hypothesis holds, the observations on the response variable y could have been observed in any order relative to the fixed value tuples in X; we need to recalculate the test statistic for each of the possible permutations of y, leaving X fixed.

A permutation test is approximated by reshuffling y k times and each time selecting a permutation randomly from the set of all possible permutations (Efron and Tibshirani, 1993). Accordingly, we calculate the approximate p-value of the test as

$$\tilde{p}\text{-value}(R^2) = \frac{\#\{\gamma = 1 \dots k \mid \hat{R}_\gamma^2 \geq \hat{R}^2\}}{k}, \tag{2}$$

where $\hat{R}_\gamma^2$ is a coefficient of determination from permutation γ of y, and $\hat{R}^2$ is a value of the test statistic for the original sample. In addition, we compute a 95% normal approximation confidence interval around the estimated p-value, which equals $\tilde{p} \pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/k}$.

The significance of each individual coefficient in model (1) is tested using the Freedman and Lane procedure (Freedman and Lane, 1983) as follows:

1) The test statistic $\hat{t}_i$ is computed from the coefficient t-statistic, when we regress y on X.
2) We compose a $n \times p$ matrix X', which contains all explanatory variables except X^i. The residuals from the regression of y on X' are subjected to permutation. For every permutation γ, we calculate a $n \times 1$ vector y' from the permuted residuals and the fixed value tuples in X'. The permutation t-statistic $(\hat{t}_i)_\gamma$ is obtained from the regression of y' on X.
3) After k replications, an approximate p-value is calculated from the generated reference distribution of t_i:

$$\tilde{p}\text{-value}(t_i) = \frac{\#\{\gamma = 1 \dots k \mid |(\hat{t}_i)_\gamma| \geq |\hat{t}_i|\}}{k}. \tag{3}$$

3.2 Implementation in 'regression-tests-1.0.0-standalone.jar'

Usage
$ java -jar regression-tests-1.0.0-standalone.jar [options] path n-replications "permutations"

Arguments:

path	Path to the csv file with an original sample
n-replications	Number of permutations replications

Options:

-t, --trace	Print stack trace
-h, --help	Print command help

Input

;;;;; 'sample-x1-x2-x3.csv'

density.asc	land_use.asc	distance_roads.	distance_indus
7.157735	9	0	0.03761
6.326746	9	0	0.032535
-1.977347	0.019694	0.154481	0.244743
5.305789	6.25	0	0.008174
...	...	...	...

The 'path' argument defines a file with sample values, e.g., 'sample-x1-x2-x3.csv'. It is expected that the first row contains variable labels. The first column should contain values of the response y.

Output

Results are saved to the 'regression-tests' folder in the root execution directory.

;;;;; 'permutation_tests.csv'

test	p-value	lower-bound-ci	upper-bound-ci
overall-test-r2	0.0001	-0.000096	0.000296
land_use.asc-test-t-stat	0	0	0
distance_roads.asc-test-t-stat	0	0	0
distance_industrial_commercial.asc-test-t-stat	0.007699	0.005986	0.009412

;;;;; 'permutation_r2_sample.csv'

value
0.850079
0.001786
0.001684
0.00143
...

4 Confidence intervals based on the percentile bootstrap

4.1 Theoretical background

Suppose that we want to estimate a confidence interval for some statistic θ in the regression model (1). Since the exact distribution of the error terms is unknown, we resample the observed data (y, X) and construct the interval using the percentile bootstrap method (Efron and Tibshirani, 1993). As a result, we carry out the following procedure:

1) The original data (y, X) is subjected to resampling with replacement. At first, a random sequence of indexes $(j_1, \dots, j_n)$ is drawn from the set $\{1, \dots, n\}$. We compose a $n \times 1$ vector y' and a $n \times (p+1)$ matrix X' by taking the selected pairs $\{(y_{j_1}, X_{j_1}), \dots, (y_{j_n}, X_{j_n})\}$. For every bootstrap replication γ, we calculate bootstrap statistic $\hat{\theta}_\gamma$ from the sample (y', X').
2) After k replications, we arrange a sequence $\hat{\theta}^*$ by taking bootstrap values $\{\hat{\theta}_\gamma\}_{\gamma=1\dots k}$ in ascending order. For the given level of confidence α, we find the $\lfloor (1-\alpha)/2\, k \rfloor$ and $\lceil (1+\alpha)/2\, k \rceil$ quantiles in $\hat{\theta}^*$ and set them as the lower and upper borders of the $100\times\ \alpha$-% percentile confidence interval respectively. Here, $\lfloor (1-\alpha)/2\, k \rfloor$ denotes the largest integer not greater than $(1-\alpha)/2\, k$ and $\lceil (1+\alpha)/2\, k \rceil$ stands for the smallest integer not less than $(1+\alpha)/2\, k$.

4.2 Implementation in 'regression-tests-1.0.0-standalone.jar'

Usage
$ java -jar regression-tests-1.0.0-standalone.jar [options] path n-replications "bootstrap-regression"

Arguments:

path	Path to the csv file with an original sample
n-replications	Number of bootstrap replications

Options:

-t, --trace	Print stack trace
-h, --help	Print command help

Input
;;;;; 'sample-x1-x2-x3.csv'

density.asc	land_use.asc	distance_roads.	distance_indus
7.157735	9	0	0.03761
6.326746	9	0	0.032535
-1.977347	0.019694	0.154481	0.244743
5.305789	6.25	0	0.008174
...	...	...	...

The 'path' argument defines a file with sample values, e.g., 'sample-x1-x2-x3.csv'. It is expected that the first row contains variable labels. The first column should contain values of the response y.

Output

Results are saved to the 'regression-tests' folder in the root execution directory.

;;;;; 'regression-stat-bootstrap.csv'

statistics	95-percent-ci-1	95-percent-ci-2	mean
land_use.asc	0.370039	0.453511	0.414008
distance_roads.asc	-45.80548	-25.320336	-34.475848
distance_industrial_commercial.asc	-5.897927	-1.483479	-3.698972
intercept	2.658128	3.310021	2.967782
r-squared	0.829551	0.872105	0.85143
mse	0.651385	0.86007	0.753846

Acknowledgments

The author would like to acknowledge DG research for funding through the FP7-funded COMPLEX project #308601, www.complex.ac.uk.

Views or opinions expressed herein do not necessarily represent those of the International Institute for Applied Systems Analysis, its National Member Organizations, or other organizations supporting the work.

References

[1] Anderson, M. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. Canadian Journal of Fisheries and Aquatic Sciences, 58(3): 626-639. DOI: 10.1139/f01-004

[2] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, 7(1): 1-26. DOI:10.1214/aos/1176344552

[3] Efron, B., & Tibshirani, R. (1993). An Introduction to the Bootstrap. New York: Chapman and Hall.

[4] Fisher, R. (1935). The Design of Experiments. Edinburgh: Oliver and Boyd.

[5] Freedman, D., & Lane, D. (1983). A Nonstochastic Interpretation of Reported Significance Levels. Journal of Business & Economic Statistics, 1(4): 292-298. DOI: 10.2307/1391660