ISSN 0737-5352-57

A SAS Macro for Theil Regression

Ann Hess
Paul Patterson
Hari Iyer
Department of Statistics
Colorado State University
Fort Collins, CO 80523

April 2002

A SAS Macro for Theil Regression

Ann Hess, Paul Patterson, Hari Iyer
Department of Statistics, Colorado State University

1. INTRODUCTION

In straight-line regression, the least squares estimator of the slope is sensitive to outliers and the associated confidence interval is affected by non-normality of the dependent variable. A simple and robust alternative to least squares regression is Theil regression, first proposed by H. Theil (1950). Theil's method actually yields an estimate of the slope of the regression line. Several approaches exist for obtaining a nonparametric estimate of the intercept. In this paper, we describe a SAS macro for implementing Theil regression where the estimation of the intercept is based on Graybill and Iyer (1994).

2. METHODOLOGY

We use β_0 and β_1 to indicate the intercept and slope of the true regression line. The input to the macro includes a bivariate dataset (with variable names) and testing level, α. The macro outputs the following quantities:

(1) estimate of slope, $\hat{\beta}_1$, of the true regression line,
(2) a p-value for testing the null hypothesis that $\beta_1 = 0$,
(3) a 1-α confidence interval for β_1,
(4) estimate of the intercept, $\hat{\beta}_0$, of the true regression line,
(5) a p-value for testing the null hypothesis that $\beta_0 = 0$, and
(6) a 1-α confidence interval for β_0.

2.1 Estimate of the Slope

Let $(x_i, y_i), i = 1, \cdots, n,$ denote the data values. Without loss of generality, assume that $x_1 \le x_2 \le \cdots \le x_n$ and $x_1 < x_n$ (i.e., there are at least two distinct *x* values in the dataset). Then let

$$N = \sum_{1 \le i < j \le n} sign(x_j - x_i),$$

where

$$sign(x) = \begin{cases} +1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

So *N* is the number of positive differences $x_j - x_i$. Now consider the set *S* of *N* distinct pairs (*i*,*j*) for which $x_j > x_i$ and define

$$S_{ij} = (y_j - y_i)/(x_j - x_i), \quad (i, j) \in S. \tag{1}$$

Thus, the S_{ij}'s are the slopes of the line segments connecting pairs of points (x_i, y_i) and (x_j, y_j) where $x_i \neq x_j$. Arrange the N quantities in (1) in ascending and denote the r^{th} smallest value of S_{ij} by $S_{(r)}$ for $r = 1,...,N$. We write $N = 2M$ if N is even or $N = 2M + 1$ if N is odd. Then the estimate of the slope is given by

$$\hat{\beta}_1 = \begin{cases} S_{(M)}, & N = 2M + 1 \\ \frac{1}{2}(S_{(M)} + S_{(M+1)}), & N = 2M. \end{cases} \tag{2}$$

In other words, the estimate of the slope is given by the median of the pairwise slopes.

2.2 Confidence Interval for the Slope

If $n \leq 10$ and there are no ties in either the independent or the dependent variable then an exact confidence interval for the slope can be computed. Confidence bounds are computed using Kendall's table (see Kendall, 1962). Given n and α, an appropriate N^* value is read from the table. Note that, because the reference distribution for computing p-values of tests is discrete, exact confidence intervals are only available for selected values of α. Hence, the value of α provided by the user will be rounded down to the nearest value given in the table. This will result in a confidence interval with coverage greater than or equal to 1-α. The actual coverage probability will be noted in the output.

Let

$$M_1 = \frac{1}{2}(N - N^*) \tag{3}$$

$$M_2 = \frac{1}{2}(N + N^*). \tag{4}$$

A confidence interval for the slope is given by the interval $(S_{(M_1)}, S_{(M_2+1)})$.

If $n \leq 10$ and ties are present in either the independent or the dependent variable then no confidence interval will be computed by the macro. A future version of the macro will accommodate this case.

If $n > 10$ (regardless of ties), then the macro computes an approximate confidence interval using the method given by Valz et al (1995). This method is based on a normal approximation whose mean and variance are appropriately adjusted for ties.

2.3 P-value for the Slope

If $n \leq 10$ then the p-value for the slope will be computed based on 1000 simulations (even when ties are present in either or both variables). If $n > 10$, then the p-value for the slope can be computed by normal approximation or simulation.

2.3.1 P-value for the Slope by Simulation

An approximate p-value for testing $\beta_1 = 0$, is obtained by simulating the reference distribution. Using the original data the following statistic is computed:

$$U^*(0) = \left\{ N\binom{n}{2} \right\}^{-1/2} \sum_{1 \le i < j \le n} sign(x_j - x_i) sign(y_j - y_i). \quad (5)$$

The following calculation is repeated K times (where K is a large integer, say 1000). Keeping the order of the *x* values in the data set fixed, the *y* values are permuted at random, resulting in a permuted dataset. Using the permuted dataset, the following statistic is computed and recorded:

$$U_k(0) = \left\{ N\binom{n}{2} \right\}^{-1/2} \sum_{1 \le i < j \le n} sign(x_j - x_i) sign(y_j - y_i). \quad (6)$$

This step is repeated K times and the value of U_k is recorded for *k* = 1,...,1000.

The estimated p-value (for a two-sided alternative) is given by counting the number of times $abs(U^*) > abs(U_k)$ and dividing by K.

2.3.2 P-value for the Slope by Normal Approximation

An approximate P-value for the slope can be obtained by using the normal approximation.

If

$$z = \frac{U^*(0)}{\sqrt{Var(U^*(0))}}, \quad (7)$$

then the p-value for the slope can be approximated by 2*(1-Prob(Z ≤ |z|)), where Z is a Normal random variable with mean 0 and variance 1.

This estimate will only be used if n>10. See Sen (1968) and Valz et al (1995) for details.

2.4 Estimate of the Intercept

The estimate of the intercept, $\hat{\beta}_0$, is computed using the method described by Graybill and Iyer (1994). The procedure is as follows.

If several *y* values are available for a given *x* value, we let y_i be their mean so we can assume that the x_i's are distinct. Arrange the *n* (distinct) observations in ascending order of *x*. If *n* is an odd number, say *n*=2*m*+1, then discard the middle observation so there are now 2*m* observations. Of course if *n* is even, no observation is discarded and *n*=2*m*. The observations are arranged as in Table 1. From Table 1, compute the quantities ***z,w,u,v*** and ***t*** which are shown in Table 2.

Table 1

Column 1	Column 2	Column 3	Column 4
y_1	x_1	y_{m+1}	x_{m+1}
y_2	x_2	y_{m+2}	x_{m+2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_m	x_m	y_{2m}	x_{2m}

Table 2

z	***w***	***u***	***v***	***t***
$z_1 = y_{m+1} - y_1$	$w_1 = x_{m+1} - x_1$	$u_1 = y_1 x_{m+1}$	$v_1 = y_{m+1} x_1$	$t_1 = u_1 - v_1$
$z_2 = y_{m+2} - y_2$	$w_2 = x_{m+2} - x_2$	$u_2 = y_2 x_{m+2}$	$v_2 = y_{m+2} x_2$	$t_2 = u_2 - v_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$z_m = y_{2m} - y_m$	$w_m = x_{2m} - x_m$	$u_m = y_m x_{2m}$	$v_m = y_{2m} x_m$	$t_m = u_m - v_m$

Compute q_i^*, where $q_i^* = t_i / w_i$ for *i*=1,...,*m*. Arrange the *m* quantities in ascending order denote the *r*th smallest value by $q_{(r)}$ for *r*=1,...,*m*. Then the estimate of the intercept is given by

$$\hat{\beta}_0 = \begin{cases} q_{(k+1)}, & m = 2k+1 \\ \frac{1}{2}(q_{(k)} + q_{(k+1)}), & m = 2k. \end{cases} \tag{8}$$

Thus, the estimate of the intercept is the median of the q_i values.

2.5 P-value for the Intercept

Under the null hypotheses $H_0 : \beta_0 = 0$, the q_i's are independently distributed, each having a Bernoulli distribution with success probability equal to 1/2. Hence a Sign Test may be used to test the above null hypothesis. The p-value for testing $\beta_0 = 0$ is computed using the built-in Sign Test in SAS.

2.6 Confidence Interval for the Intercept

A confidence interval for the intercept is computed by finding a confidence interval for the median of the q_i values. This is done using a built in SAS function, and is based on the Sign Test.

3. MACRO INSTRUCTIONS

Begin by downloading the **Theil.sas** file from the website:
www.stat.colostate.edu/~hari/nps.

Then create a SAS dataset using a data step or by importing a file.

Next, use **filename** and **%include** statements to indicate the name and location of the Theil.sas file. For example if Theil.sas was saved to a disk on the A: drive, the following commands would be used:

```
filename Theil 'A:Theil.sas';
%include Theil;
```

To run the program, edit the following command:

```
%SenSlope(data,x,y,alpha,nsim,seed);
```

where *data* should be replaced by the name of your SAS dataset, *x* should be replaced by the name of the independent variable (as given in the SAS dataset), *y* should be replaced by the name of the dependent variable (as given in the SAS dataset) The value of *alpha* should be a number between 0 and 0.5 (the choice *alpha* = 0.05 will yield 95% confidence intervals). The value of *nsim* should be a number greater than or equal to 0. If the dataset contains 10 or fewer observations, the p-value for the slope will be computed based on 1000 simulations regardless of the value of *nsim*. However, if there are more than 10 observations then if *nsim*=0 the p-value for the slope will be based on the normal appoximation and if *nsim*>0 the p-value for the slope will be based on a simulation with *nsim* equal to the number of simulations used (see Section 2.3 for details). The value of *seed* should a positive integer between 1 and 2^{23}-1 (if required, this will be the seed for the simulation).

The program will output estimates, confidence intervals, and p-values for both the slope and the intercept.

Notes:

- If the dataset contains more than 10 observations and *nsim*>0 the p-value for the slope will be calculated by simulation and will depend on the value *nsim*. So, changing the value of *nsim* and/or *seed* will usually result in a change in the estimate of the p-value for the slope.
- The running time of the macro depends on the value of *nsim* (as well as the size of the dataset). Values of *nsim*>1000 may result in ***very*** long run times.
- The macro will produce a file **c:\theil_junk** which contains extraneous SAS output and should be disregarded. It can be deleted at any time.
- When the p-value for the slope is computed by simulation, the following additional output will be given: the value of *nsim*, a confidence interval for the p-value, and the value of *seed*.

4. EXAMPLE

The following example is taken from Graybill and Iyer (1994). Suppose an investigator is studying the association between sulfur dioxide (SO2) concentrations in a national park and the rate of emission of SO2 by a coal burning power plant 25 miles away. To assess the power plant's SO2 contribution to the national park, recordings were made of X, the SO2 output by the plant in tons/hour, as well as Y, the SO2 concentrations at the national park in micrograms/cubic meter. The investigator would like a straight line regression equation relating Y to X using Theil's method. (One should verify using a scatter plot that a straight line model is reasonable). The macro input is given in the file **so2example.sas**.

MACRO INPUT:

```
data so2;
input y x @@;
datalines;
 5.21 1.92
 7.36 3.92
16.26 6.80
10.10 6.32
 5.80 2.00
 8.06 4.32
 4.76 2.40
 6.93 2.96
 9.36 3.52
10.90 4.24
12.48 5.12
11.70 5.84
 7.44 3.60
 6.99 2.80
;

filename Sen 'A:Theil.sas';
%include Sen;

%SenSlope(so2,x,y,0.10,1000,259);
```

MACRO OUTPUT:

```
Quantity      Estimate       LB          UB      Coverage    pValue

Intercept     1.246429    0.36414    8.42971     92.9688    0.1250
Slope         1.750000    1.02500    2.27679     90.0000    <.0001

The p-value for the Slope is based on 1000 simulations.
The 90 % CI for the p-value is ( 0 , 0.0029912 ).
The seed for the simulation was 259 .
```

In the output, LB stands for “lower bound” and UB stands for “upper bound” for a confidence interval on the appropriate parameter (see the column labeled “Quantity”). The actual coverage probabilities of the confidence intervals may differ from the nominal coverage probability requested (due to the discreteness of the reference distributions). Hence, the actual coverage probability is listed under the column labeled “Coverage”.

5. AUTHOR CONTACT

Questions and Comments about this macro should be addressed to Ann Hess at
hess@stat.colostate.edu.

6. REFERENCES

Graybill, F.A. and Iyer, H.K., *Regression Analysis : Concepts and Applications*, Duxbury Press: Belmont, CA, 1994.

Kendall, M.G., *Rank Correlation Methods*, Charles Griffin and Company: London, 1962.

Sen, P.K., "Estimates of the Regression Coefficient Based on Kendall's Tau", *American Statistical Journal*, 63(324), 1968.

Theil, H., "A rank-invariant method of linear and polynomial regression analysis" I, II and III, *Nederl. Akad. Wetensch. Proc*., 53, 1950.

Valz, P.D., McLeod, A.I., Thompson, M.E., "Cumulant Generating Function and Tail Probability Approximations for Kendall's Score with Tied Rankings", *Annals of Statistics*, 23(1), 1995.