Chapter 314

Robust Linear Regression (Passing-Bablok Median-Slope)

Introduction

This procedure performs robust linear regression estimation using the Passing-Bablok (1988) median-slope algorithm. Their original algorithm (1983, 1984) was designed for method comparisons in which it was desired to test whether the intercept is zero and the slope is one. This procedure extends that algorithm for the case where the main interest is in estimating the intercept and slope in the linear equation

$$Y = \beta_0 + \beta_1 X.$$

The estimate of the slope ($B1$) is calculated as the median of all slopes that can be formed from all possible pairs of data points, except those pairs that result in a slope of 0/0. To correct for estimation bias, the median is *shifted* by a factor K which is one-half the number of slopes that are less than zero. This creates an approximately unbiased estimator.

The estimate of the intercept ($B0$) is the median of $\{Y_i - B1\ X_i\}$.

This procedure is used for *transference*: when you want to rescale one reference interval to another scale.

Experimental Design

Typical designs suitable for Passing-Bablok regression include up to 3000 paired measurements, (x_i, y_i), $i = 1, \ldots, n$, similar to the common input for simple linear regression. Typical data of this type are shown in the table below.

Typical Data for Passing-Bablok Median-Slope Regression

Subject	X	Y
1	7	7.9
2	8.3	8.2
3	10.5	9.6
4	9	9
5	5.1	6.5
6	8.2	7.3
7	10.2	10.2
8	10.3	10.6
9	7.1	6.3
10	5.9	5.2

© NCSS, LLC. All Rights Reserved.

Technical Details

The methods and results in this chapter are based on the formulas given in Passing and Bablok (1983, 1984) and Bablok, Passing, Bender, and Schneider (1988).

Assumptions

Passing-Bablok regression requires the following assumptions:

1. The relationship between *X* and *Y* is linear (straight-line).
2. No special assumptions are made about the distributions (including the variances) of *X* and *Y*.

Passing-Bablok Estimation

Define X_i and Y_i, $i = 1, \ldots, n$, as the values for two variables, each sampled with error to give the observed values x_i and y_i, respectively. For each of the $N = \binom{n}{2}$ possible pairs of points, define the slope by

$$S_{ij} = \frac{y_i - y_j}{x_i - x_j}$$

with the following substitutions

If $x_i = x_j$ and $y_i = y_j$, the result is 0/0 which is undefined. Omit these pairs from the calculation of the slope.

If $x_i = x_j$ and $y_i > y_j$, $S_{ij} = +\infty$.

If $x_i = x_j$ and $y_i < y_j$, $S_{ij} = -\infty$.

If $x_i > x_j$ and $y_i = y_j$, $S_{ij} = +0$.

If $x_i < x_j$ and $y_i = y_j$, $S_{ij} = -0$.

The value of *K* is computed as [*neg*/2] where *neg* is the number of S_{ij}'s that are negative (including -0's and -∞'s).

The slope *B*1 is the shifted median of *Sij*, where the median is shifted to the right *K* steps.

Using this slope, calculate intercept *B*0 as the median of all *n* of the quantities $y_i - B1x_i$.

Confidence Bounds

Calculate the confidence bounds for β_1 as follows. Let $z_{\alpha/2}$ be the 1 – α/2 quantile of the standard normal distribution. Let

$$C_{\alpha/2} = z_{\alpha/2}\sqrt{\frac{n(n-1)(2n+5)}{18}}$$

Now calculate

$$M_1 = \left[\frac{N - C_{\alpha/2}}{2}\right]$$

Here, [U] rounds U to the nearest integer.

Also calculate

$$M_2 = N - M_1 + 1$$

© NCSS, LLC. All Rights Reserved.

Finally, a confidence interval for β_1 is given by

$$S_{(M_1+K)} \leq \beta_1 \leq S_{(M_2+K)}$$

Where the *Sij's* are sorted.

Confidence limits for the intercept are calculated as follows. Let $B1_L$ and $B1_U$ be the lower and upper limits for the slope from the last calculation. Now calculate the limits for the intercept as

$$B0_L = \text{median}\{y_i - B1_U x_i\}$$

and

$$B0_U = \text{median}\{y_i - B1_L x_i\}$$

Negative Kendall's Tau

If the data exhibit a negative value of Kendall's tau, the substitution *w* = -*y* is used was switch the correlation to a positive value. Once estimation is complete, the estimate of *B*1 is set to $-B1_w$.

Kendall's Tau Test of the High Correlation Assumption

Passing and Bablok (1983) recommended that a preliminary two-sided test be conducted to determine if Kendall's tau correlation between X and Y is significantly different from zero. They also indicate that this correlation must be positive. Kendall's tau correlation is well documented in the *Correlation* procedure.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables and estimation parameters used in the analysis.

Variables

Y Variable

Specify a single data column to be used for the Y (dependent) variable. This column must contain continuous values.

Robust estimates of B0 and B1 will be found using the Passing-Bablok median-slope algorithm. The estimated linear equation will be of the form

Y = B0 + B1 × X

You can enter the column name or number directly, or click the button on the right to display a Column Selection window that will let you select the column from a list.

X Variable

Specify a single data column to be used for the X (independent) variable. This column must contain continuous values.

Robust estimates of B0 and B1 will be found using the Passing-Bablok median-slope algorithm. The estimated linear equation will be of the form

Y = B0 + B1 × X

You can enter the column name or number directly, or click the button on the right to display a Column Selection window that will let you select the column from a list.

© NCSS, LLC. All Rights Reserved.

Grouping Variable (Optional)

Grouping Variable

Enter a single categorical grouping variable. The values of this variable indicate which category each subject belongs in. Values may be text or numeric. The grouping variable is optional.

A separate analysis will be performed for each group.

Reports Tab

This tab controls which statistical reports are displayed in the output.

Significance Levels and Confidence Levels

Confidence Level

This confidence level, entered as a percentage, is used in the calculation of confidence intervals for regression coefficients.

Typical confidence levels are 90%, 95%, and 99%, with 95% being the most common.

Range

50 < Confidence Level < 100

Assumptions Alpha

Specify the alpha value (significance level) used for tests of assumptions. Alpha is the probability of rejecting the null hypothesis when it is actually true.

Recommended

Often, an alpha of 0.05 is used. However, many statisticians recommend a higher value of alpha for tests of assumptions such as 0.10 or even 0.20.

Range

Typical choices for alpha range between 0.001 and 0.200.

Select Reports

Run Summary to Residuals

Each of these options specifies whether the indicated report is displayed.

Report Options

The following options control the format of the reports.

Variable Names

Specify whether to use variable names, variable labels, or both to label output reports. In this discussion, the variables are the columns of the data table.

- **Names**
  Variable names are the column headings that appear on the data table. They may be modified by clicking the Column Info button on the Data window or by clicking the right mouse button while the mouse is pointing to the column heading.

© NCSS, LLC. All Rights Reserved.

- **Labels**

  This refers to the optional labels that may be specified for each column. Clicking the Column Info button on the Data window allows you to enter them.

- **Both**

  Both the variable names and labels are displayed.

Comments

1. Most reports are formatted to receive about 12 characters for variable names.
2. Variable Names cannot contain blanks or math symbols (like + - * / . ,), but variable labels can.

Value Labels

Value Labels are used to make reports more legible by assigning meaningful labels to numbers and codes.

The options are

- **Data Values**

  All data are displayed in their original format, regardless of whether a value label has been set or not.

- **Value Labels**

  All values of variables that have a value label variable designated are converted to their corresponding value label when they are output. This does not modify their value during computation.

- **Both**

  Both data value and value label are displayed.

Example

A variable named GENDER (used as a grouping variable) contains 1's and 2's. By specifying a value label for GENDER, the report can display "Male" instead of 1 and "Female" instead of 2. This option specifies whether (and how) to use the value labels.

Report Options – Decimal Places

Item Decimal Places

These decimal options allow the user to specify the number of decimal places for items in the output. Your choice here will not affect calculations; it will only affect the format of the output.

- **Auto**

  If one of the "Auto" options is selected, the ending zero digits are not shown. For example, if "Auto (0 to 7)" is chosen,

  0.0500 is displayed as 0.05
  1.314583689 is displayed as 1.314584

  The output formatting system is not designed to accommodate "Auto (0 to 13)", and if chosen, this will likely lead to lines that run on to a second line. This option is included, however, for the rare case when a very large number of decimals is needed.

© NCSS, LLC. All Rights Reserved.

Plots Tab

These options specify which plots are produced as well as the plot format.

Passing-Bablok Regression Plots

Passing-Bablok Regression Scatter Plot

Indicate whether to display this plot. Click the plot format button to change the plot settings.

Show Combined Plot (If Grouping Variable Present)

If you have a grouping variable present, this option allows you to plot all groups on one plot for comparison. Passing-Bablok regression is still performed on each group separately.

Residual Plot

Residual Plot

Indicate whether to display this plot. Click the plot format button to change the plot settings.

© NCSS, LLC. All Rights Reserved.

Example 1 – Passing-Bablok Median-Slope Regression

This section presents an example of how to run a Passing-Bablok regression analysis of the data in the *PassBablok1* dataset. This dataset contains measurements from two measurement methods on each of 30 items. The goal is to find robust estimates of the coefficients in the linear regression equation. You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the procedure window.

1 Open the PassBablok 1 example dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **PassBablok 1.NCSS**.
- Click **Open**.

2 Open the Robust Linear Regression (Passing-Bablok Median-Slope) procedure window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Robust Linear Regression (Passing-Bablok Median-Slope)** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- Select the **Variables** tab.
- For **Y Variable**, enter **Method3**.
- For **X Variable**, enter **Method1**.
- Leave the **Grouping Variable** blank.

4 Specify the reports.

- Select the **Reports** tab.
- In addition to the reports already selected, check **Residuals**.

5 Specify the plots.

- Select the **Plots** tab.
- Click on the **Passing-Bablok Regression Scatter Plot** button.
- Click on the **Residuals Plot** button.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Run Summary Report

Run Summary

Item	Value	Item	Value
Y Variable	Method3	Rows Used	30
X Variable	Method1	B0: Intercept	99.9907
		B1: Slope	-0.9983

This report gives a summary of the input and various descriptive measures about the Passing-Bablok regression.

© NCSS, LLC. All Rights Reserved.

Descriptive Statistics Report

Item	Y	X
Variable	Method3	Method1
N	30	30
Mean	37.00667	63.14
Std Dev	23.5886	23.6153
Std Error	4.3067	4.3115
COV	0.6374	0.3740
Minimum	1.4	14.2
First Quartile	16.925	48.8
Median	35.85	61.4
Third Quartile	52.575	84.1
Maximum	87.3	99.1

This report gives descriptive statistics about the variables used in the regression.

Kendall's Tau Correlation Confidence Interval and Hypothesis Test

Hypothesis Test of Y = X

Kendall's Tau Correlation	Lower 95%: Conf. Limit of Tau	Upper 95%: Conf. Limit of Tau	Z-Value for Testing H0: ρ = 0	P-Value	Reject H0 that ρ = 0 at α = 0.05?
-0.9724	-0.9833	-0.9546	-7.5289	0.0000	Yes

This procedure assumes that Kendall's Tau is significantly different from zero.
This assumption cannot be rejected.

The section reports an analysis of Kendall's tau correlation. It provides both a confidence interval and a significance test. The main thing to look for here is that the absolute correlation (-0.9724 in this example) is large. This can be surmised from both the confidence interval and the p-value.

Regression Coefficient

Item	Intercept B0	Slope B1
Regression Coefficient	99.9907	-0.9983
Lower 95% Conf. Limit of β(i)	99.1920	-1.0098
Upper 95% Conf. Limit of β(i)	100.7432	-0.9854

This section reports the regression coefficients, along with their analytic confidence limits. **These estimates are the main focus of the analysis**.

© NCSS, LLC. All Rights Reserved.

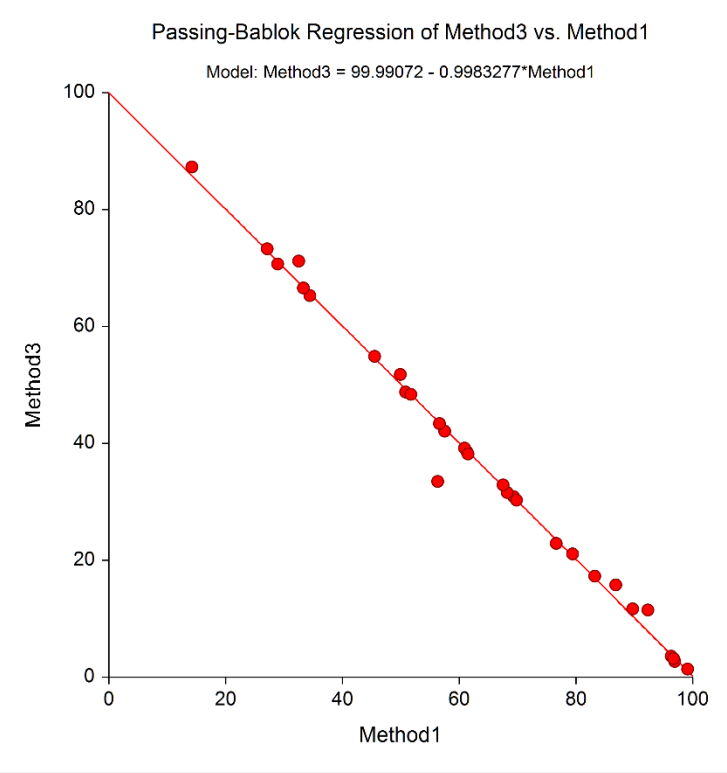
Residuals Report

Residuals

Row	X	Y	Difference Y - X	Predicted Yhat	Residual (Y-Yhat)
1	69.3	30.9	-38.4000	30.8066	-0.0934
2	27.1	73.3	46.2000	72.9360	-0.3640
3	61.3	38.6	-22.7000	38.7932	0.1932
4	50.8	48.8	-2.0000	49.2757	0.4757
5	34.4	65.3	30.9000	65.6482	0.3482
6	92.3	11.5	-80.8000	7.8451	-3.6549
7	57.5	42.1	-15.4000	42.5869	0.4869
8	45.5	54.9	9.4000	54.5668	-0.3332

This section reports the residuals and the difference for each of the input data points.

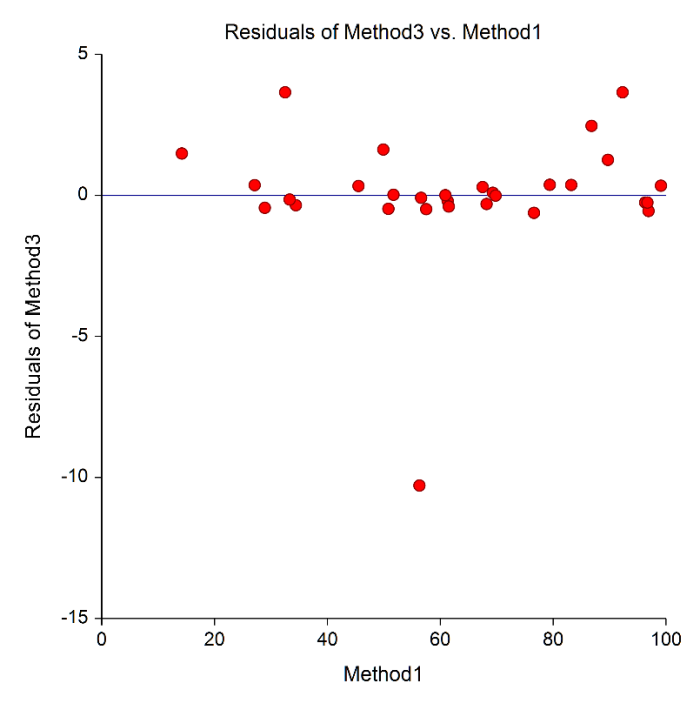
Passing-Bablok Regression Scatter Plot



This report shows the fitted Passing-Bablok regression line.

© NCSS, LLC. All Rights Reserved.

Residual Plot



This plot emphasizes the deviation of the points from the regression line.

© NCSS, LLC. All Rights Reserved.