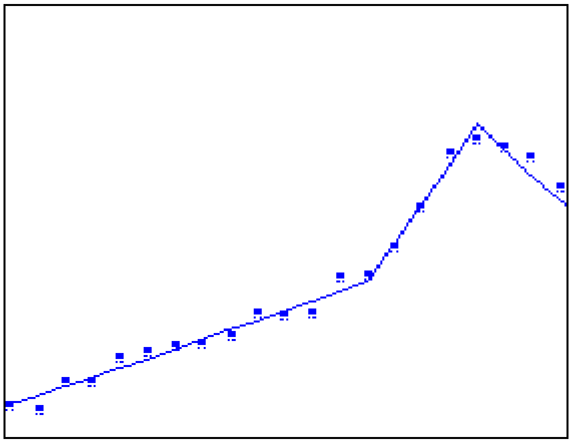
JoinPoint User’s Guide

Getting Started

The JoinPoint Regression Program is a Windows-based statistical software package that analyzes JoinPoint models. The software enables the user to test whether or not an apparent change in trend is statistically significant.

JoinPoint fits the selected trend data (e.g. cancer rates) into the simplest JoinPoint model that the data allow. The resulting graph is like the figure below, where several different lines are connected together at the "JoinPoints".



JoinPoint also allows the user to view one graph for each JoinPoint model, from the model with the minimum number of JoinPoints to the model with maximum number of JoinPoints.

You may use any software to create data files to be analyzed by JoinPoint (see Input Data File). The SEER*Stat software provides an easy mechanism for creating data files containing cancer rates or counts that can be analyzed by JoinPoint. The controls on the JoinPoint **Specifications** tab will be set automatically if you load data files exported from SEER*Stat. To learn more about the SEER*Stat software visit the SEER*Stat Web site at:

http://seer.cancer.gov/seerstat/

More:
System Requirements
Citation
Technical Support
Specifications Tab

Citation

The JoinPoint regression model and permutation tests for identifying changes in trend are described in:

Methods Citation:

Kim HJ, Fay MP, Feuer EJ, Midthune DN. "Permutation tests for JoinPoint regression with applications to cancer rates" *Statistics in Medicine* 2000; 19:335-351: (correction: 2001;20:655).

The present document describes the use of the JoinPoint Regression Program for Windows.

Software Citation:

JoinPoint Regression Program, Version 3.5 - April 2011; Statistical Methodology and Applications Branch and Data Modeling Branch, Surveillance Research Program National Cancer Institute.

For more information concerning the statistical calculations used by the JoinPoint program see Statistical Notes.

Technical Support

For an up-to-date list of known problems, frequently asked questions (FAQs) and resolutions, please consult the JoinPoint Web site at:

http://surveillance.cancer.gov/JoinPoint/faq/

For technical support beyond the help system or Web site, please send email to:

JoinPoint@imsweb.com

System Requirements

The system requirements to run JoinPoint are a personal computer with at least:

- a Pentium or equivalent processor
- 32 MB application RAM
- 2 MB hard disk space
- a 32-bit version of Microsoft Windows (Windows 95 or later)
- screen resolution set to 800 by 600 pixels or greater

Frequently Asked Questions

For more help using JoinPoint check the Surveillance and Research Program Web site for JoinPoint: Frequently Asked Questions at http://surveillance.cancer.gov/JoinPoint/faq/, or look in the back of this User's Guide in the Frequently Asked Questions section.

Setting the JoinPoint Parameters

When using JoinPoint, work on each tab in sequence - from left to right, and from top to bottom within each tab.

Specifications Tab

On the **Specifications** tab you will describe the data file to be analyzed by the JoinPoint program and select the options for the modeling of the data.

Advanced Tab

Use the **Advanced** tab to specify the modeling method, the criteria used to determine the locations of the JoinPoints, choose the model selection method, and enter the permutation options – the significance level and the number of permutations.

You can save all your settings in a JoinPoint session file (called a JoinPoint parameter file in version 2.7 and earlier) by selecting **File > Save As...** from the top of the main window and choosing a name for the session file.

When you have made all of your selections on the JoinPoint tabs, click ⚡ to execute the JoinPoint session.

Note, the session file is saved as part of the output file, so you can select **Output > Retrieve Session** to retrieve the session that was used to create the output, even if that session file wasn't specifically saved.

Specifications Tab

The purpose of the Specifications tab is to set up your regression analysis by identifying the input data file, selecting the model type, selecting and defining the dependent, independent, and any by variables, selecting the heteroscedastic errors option, and setting the minimum and maximum possible number of JoinPoints.

More:

Input Data File
Type of Model
Independent Variable
By-variables
Heteroscedastic Errors Option
Number of JoinPoints
Test for Pairwise Differences

Input Data File

The input data file contains a dependent variable and an independent variable or covariate. The dependent variable may be an age-adjusted rate, crude rate, or count. The independent variable is typically the year. The data file may also contain standard errors (for rates), a population or offset variable (for counts) and by-variables such as age group, sex, or race. These variables do not have to be in any particular order. You identify the dependent variable, etc., by specifying its name or column position in the data file.

SEER*Stat provides an easy mechanism for creating JoinPoint data files and setting the controls on the JoinPoint **Specifications tab**. In SEER*Stat, select "Matrix | Export" when viewing the SEER*Stat matrix to create a file in the format of a JoinPoint Input Data File. In JoinPoint, select "New Session" from the **File** menu and then select "SEER*Stat Export Dictionary" in the "Create New Session" window. The input data file name, variable locations and file format information will be loaded from the SEER*Stat Export Dictionary.

The Heteroscedastic Errors Option is automatically set from the SEER*Stat Export Dictionary, depending on whether or not the standard errors are included. If they are not included, Constant Variance is selected. If they are included, Input Standard Error of Dependent Variable is selected.

More:
Format of the JoinPoint Input Data File
Using SEER*Stat to Create JoinPoint Input Data Files
Independent Variable
By-variables
Specifications Tab

Format of the JoinPoint Input Data File

Each record in the data file has the following layout:
Var(1) Var(2) . . . . Var (k)

The variables may be separated by a single character. This field delimiter may be a tab, space, comma, or semi-colon.

The records must be sorted by: by-variables, independent variable.
Missing values may be represented by either a space or a period.

Using SEER*Stat to Create JoinPoint Input Data Files

JoinPoint requires a specific file format for the input data file. Therefore, the following export options *must* be used in SEER*Stat when creating data files to be used in JoinPoint:

> The data file must be a text file (use a .txt extension or a gzipped text file (use a .gz extension)
> The Output Variables option must be set to "Numeric Representation"
> The Line Delimiter option must be set to "DOS/Windows (CR/LF)"
> The Field Delimiter may be any one of the choices. However, if commas are used as the field delimiter then the option to "Remove All Thousand Separators (Commas)" *must* be checked.
> "Remove Flags (Footnote Characters)" must be selected.

The independent variable (year) must be an individual value. Therefore, be sure to create a user-defined variable in SEER*Stat that does *not* contain any ranges. For example, when creating a new "Year of Diagnosis" variable remove the range variable (1973-2004) for all years combined.

The records for each by-group must be contiguous. If N = number of records for the first by-group, lines 1 through N must contain all values for the first by-group. Within each by-group, the records must be sorted by the independent variable (this is typically year).

To be sure that the sort order is correct, on the **SEER*Stat Table** tab, put all table variables in the row dimension. Make the independent variable (year) the last variable in the list of row variables. Alternatively, you could set the independent variable as the column variable.

The SEER Cancer Statistics Review (CSR) uses the same JoinPoint Regression Program. Their analysis uses the "Input Standard Error of Dependent Variable" setting for the Heteroscedastic Errors Option with the standard error of the rate used as an estimate of the standard deviation. Using these options, and the current default number of permutations (4499), if the JoinPoint Regression Program chooses the model with 0 JoinPoints, the annual percentage rate change will agree with the calculation of this value given in the CSR.

Note: To obtain the SEER*Stat Program visit the SEER web page at:

http://seer.cancer.gov/seerstat/

Type of Model

Select a linear or log-linear model.

- Linear Model: y = x'beta + e
- Log-linear Model: ln(y) = x'beta + e

By-variables

A separate analysis will be performed for each by-group. The by-variable information is automatically loaded when you open the SEER*Stat Export dictionary associated with your data file. If you created your data file with software other than SEER*Stat, you must manually add by-variables. To add a by-variable, click the Add... button, select the desired variable, and click OK.

More:
Format of the JoinPoint Input Data File
Edit Variable Window

Edit Variable Window

The following attributes must be set for each by-variable:

Variable name:

A maximum of 40 characters may be used for the name of the variable.

Position:

The field position of the variable in the input data file is automatically filled in.

Coding Information (Value = Label):

The coding information for the variable is displayed in a text box containing one valid value, an equal sign, and a label in each line. For example:

1=Brown

2=Blue

3=Green

Valid values must include all codes used in the data file for this variable.

Labels may have any meaningful text including numbers, letters, and special characters. They must also be a reasonable length.

Independent Variable

The independent variable is defined on the **Specifications** tab by its name or column position. These controls will automatically be set when the SEER*Stat Export Dictionary is loaded. If you created your data file with software other than SEER*Stat you must set these controls.

The independent variable or covariate is typically the year of diagnosis or year of death. JoinPoint also allows the independent variable to have non-integer or negative values. You are limited to a 40-character label for the independent variable name.

The "Shift Data Points by" option allows all the values for the independent variable to be shifted up by a fixed value. This is done by simply entering a value for "Shift Data Points by". For example, if your independent variable is years (input as 1975, 1976,...), but you would like these points to represented on the graph at the midpoint of the years (1975.5, 1976.5, ...), then you would enter the value 0.5 for this option. *Shifting the data points will change the location of the JoinPoints and the intercepts but will not change the slopes or APCs*.

The effect of entering a value for this option can be seen by clicking on "Define" for the Independent Variable. Note: this value cannot be greater than the maximum interval between data points.

One reason for employing this option would be because each data point represents a summary of data collected over a time interval. For example, cancer incidence or mortality data is often collected over the course of a year, and is usually entered as a whole year value, e.g. 1990, 1991. Instead, one may want to shift by half a year so that the data point is represented as the midpoint of the interval, e.g. 1990.5, 1991.5). This is especially important if JoinPoints are allowed to occur at places other than the data points (either in continuous time using Hudson's algorithm, or using a grid search where grid points are allowed between data points). If the data points are not shifted, the results may be counterintuitive. For example, without a shift, a JoinPoint located at 1989.50 represents the beginning of 1990, and 1990.49 represents the end of 1990. If the points are shifted by half a year, then 1990.00 represents the start of the year and 1990.99 represents the end of the year.

Heteroscedastic Errors Option

Select one of the four Heteroscedastic Errors Options. You must then specify the variable required for that option.

The error random variable in a model is homoscedastic if the variance of the error is constant; otherwise, the error is heteroscedastic. Often the homoscedastic assumption is violated, particularly when the variance of the error varies with time. This option allows the user to choose between a model where errors are assumed to have constant variance (homoscedasticity) and a model where nonconstant variance (heteroscedasticity) is assumed.

Heteroscedasticity is handled by JoinPoint using weighted least squares (WLS). The weights in WLS are the reciprocal of the variance and can be specified in multiple ways, three of which are implemented here. The first, “input standard error of the dependent variable” allows the user full flexibility to specify the standard error at each time period. These standard errors can come from the output of SEERStat. The last two options, “Poisson variance”, estimate the nonconstant variance by assuming the dependent variable counts follow a Poisson distribution. If “Poisson Variance using Crude Rate” is selected, a population parameter needs to be entered in order to convert the crude rate to a count.

More:
Constant Variance
Input Standard Error of Dependent Variable
Poisson Variance Using Crude Rate
Poisson Variance Using Count

Constant Variance

This selection assumes the random errors in the regression model are homoscedastic (have constant variance) and estimates the regression coefficients by ordinary least squares for both model ln (y) = xb and model y = xb.

More:
Input Standard Error of Dependent Variable
Poisson Variance Using Crude Rate
Poisson Variance Using Count

Input Standard Error of Dependent Variable

This is the default selection. It assumes that the random errors are heteroscedastic (have non-constant variance) and estimates the regression coefficients by weighted least squares, where weights at each point are:

- For model y = xb

w = 1/v, where v is the square of the std dev that has been input for that point.

- For model ln (y) = xb

w = (y^2)/v, where y^2 is the square of the response for that point and v is the square of the std dev that has been input for that point. (Motivated by delta method.)

More:
Constant Variance
Poisson Varriance Using Crude Rate
Poisson Variance Using Count

Poisson Variance Using Crude Rate

This selection assumes the response variable is y = c/p, where c is the adjusted count, which equals either:

- The count, if there are no counts equal to zero.
- The count plus one half, if there exist some counts equal to zero, and p is the adjusted population similarly defined. In other words one half is added to the populations whenever one half is added to the counts.

Assume the random errors for c are Poisson, and estimate the regression coefficients by weighted least squares, where the weights at each point are:

- For model y = xb

  $w = p^2/c$, where c is the adjusted count, and p^2 is the square of the adjusted population for that point.

- For model ln(y) = xb

  w = c, where c is the adjusted count for that point. (Motivated by delta method.)

More:
Constant Variance
Input Standard Error of Dependent Variable
Poisson Variance Using Count

Poisson Variance Using Count

This selection assumes the response variable is y = c, where c is the adjusted count, which equals either:

- The count if there are no counts equal to zero
- The count plus one half if there exist some counts equal to zero.

Assume the random errors are Poisson, and estimate the regression coefficients by weighted least squares, where weights at each point are:

- For model y = xb
  w = 1/c, where c is the adjusted count for that point.
- For model ln(y) = xb
  w = c, where c is the adjusted count for that point. (Motivated by delta method.)

More:
Constant Variance
Input Standard Error of Dependent Variable
Poisson Variance Using Crude Rate

Number of JoinPoints

Enter the minimum and maximum number of JoinPoints to fit where:

0 $\leq$ min $\leq$ max $\leq$ 9 when the Grid Search method is selected.

-or-

0 $\leq$ min $\leq$ max $\leq$ 4 when Hudson's method is selected.

As of Version 3.5, the default value for the maximum number of JoinPoints depends on the number of data points. This value can be changed by the user, subject to having a minimum number of data points necessary to satisfy two pre-set rules:

- A JoinPoint cannot occur within a user-specified number (default: 3) of data points from the beginning or end of a series.
- There must be at least a user-specified number (default: 4) of data points between two JoinPoints.

The default maximum number of JoinPoints is a recommendation based on the same metrics used to determine the minimum number of data points (or conversely the maximum number of JoinPoints for a given number of data points). The default is based on the following recommendations:

- At least seven data points should be observed in order to consider allowing a JoinPoint.
- There should be, on average, at least four data points between consecutive JoinPoints.

These algorithmic recommendations lead to the following default maximum number of JoinPoints.

Number of Data Points	Default Maximum Number of JoinPoints
0 - 6	0
7 -11	1
12 -16	2
17 - 21	3
22 -26	4
27+	5

Note: Due to computational intensity, the default for the maximum number of JoinPoints is capped at 5 when the Grid Search Method is selected and is capped at 4 when the Hudson’s Method is selected. The Grid Search allows a maximum up to 9, but those runs could take quite a long time to complete. The maximum for Hudson’s is 4.

More:
Number of Observations

Test for Pairwise Differences

The option to test for pairwise differences between by-groups runs a statistical test to compare whether two sets of data are parallel or coincident. All 2-way combinations of the innermost by-variable are tested.

The main goal of the comparability test is to compare two sets of trend data whose mean functions are represented by JoinPoint regression. Specific interests are on testing (i) whether two JoinPoint regression functions are identical (test of coincidence) or (ii) whether the two regression mean functions are parallel (test of parallelism), and the details can be found in Kim et al. (2004):

> H. J. Kim, M. P. Fay, B. Yu, M. J. Barrett and E. J. Feuer (2004), Comparability of Segmented Line Regression Models, Biometrics, 1005-1014.

Note: this could add substantially to the run time. Also, Autocorrelated Errors Options are not available with this test.

For more details, see:

http://srab.cancer.gov/JoinPoint/comparabilitytest.html

Advanced Tab

The purpose of the **Advanced** tab is to specify the modeling method (Grid Search or Hudson's), autocorrelated errors options, constraints on the location(s) of the JoinPoints, and. the model selection method (Permutation Test, BIC, or Modified BIC). Also, the permutation test options (the significance level and the number of permutations) are set here.

More:
Method
Autocorrelated Errors Option
Points Adjacent to the JoinPoints
Model Selection Method

Method – Grid Search or Hudson's

JoinPoint allows two different methods for model fitting – Grid Search or Hudson's. The Grid Search has a discrete number of locations that are testing to find the best model fit while Hudson's allows for continuous testing. Here are the details for the two different methods:

More:
Grid Search Method
Hudson's Method

Grid Search Method – Details

Prior to Version 3.1, JoinPoint always used the Grid Search Method to determine the best fit for each individual model. With the default settings, this only allows the JoinPoints to occur exactly at an observation. This does not, however, find the best fit. A better fit can be achieved by using a finer grid – by changing the setting for "Number of points to place between adjacent observed *x* values in the grid search" to something larger than the default of zero. So, the Grid Search Method, creates a "grid" of all possible locations for JoinPoints specified by the settings, and tests the SSE at each one to find the best possible fit. With lower values for "Number of points to place between...", this method is computationally more efficient.

More:
Hudson's Method
Number of Observations

Hudson's Method – Details

Prior to Version 3.1, the JoinPoint program always used the Grid Search Method to determine the best fit for each individual model. Hudson's Method does a continuous testing between observed *x* values to find the best model, so the fit will be better than even a fine grid of 9 for "Number of points to place between adjacent observed *x* values in the grid search". Hudson's Method is still much more computationally intensive than the Grid Search with zero "points between", so the execution time required by the calculation engine may excessive. But it is faster than a very fine grid while also achieving a better fit.

Note, since the fit is better for each model and the SSEs are lower, this can impact which JoinPoint model is selected as the best one. The general tendency is that it selects a model with fewer JoinPoints than the Grid Search since the SSE can be substantially lower for that model with Hudson's Method.

More:
Grid Search Method
Points Adjacent to the JoinPoint (Number of Observations)

Autocorrelated Errors Option

If you select "Fit an uncorrelated errors model" the program assumes the random errors in the regression model are uncorrelated and estimates the regression coefficients by ordinary least squares (unless the errors are heteroscedastic; see Heteroscedastic Errors Option).

If you select "Fit an autocorrelated errors model based on the data", the autocorrelation parameter will be computed separately for each by-group. Details on how this parameter is computed will be posted on the Web site.

If you select "Fit an autocorrelated errors model with parameter =", you must input an autocorrelation parameter (usually between zero and one) which represents the correlation between adjacent points. The program then assumes the random errors are autocorrelated and estimates the regression coefficients by weighted least squares. The autocorrelation model assumes corr(ei,ej) = phi**|i-j|, where ei and ej are the errors corresponding to the ith and jth points and phi is the autocorrelation parameter chosen. This option makes sense only with equally spaced points.

Although the autocorrelation may be estimated from the data (see Kim, et al., Statistics in Medicine, 2000, 335-351), correcting for autocorrelation with this estimate may seriously reduce the power to detect JoinPoints. (See section 2.3). We found in our simulations on table IV of that paper that adjusting for autocorrelation was helpful in maintaining proper size of the tests of JoinPoints when there was large autocorrelation. We also found that if there was no autocorrelation that the adjustment seriously affected the power of the test to detect JoinPoints. For example we see in Table IV (b) with phi = 0, the power goes from 90% to 68 %. This is because it is difficult to differentiate between autocorrelation and JoinPoints in a model.

If you suspect that your data is positively autocorrelated, we suggest using the "Fit an autocorrelated errors model with parameter" option to see how sensitive your results are to changes in autocorrelation. The option should be used is as follows:

1. Fit the model with the uncorrelated errors option.
2. If the user suspects that there is positive autocorrelation in the data, then repeat the analysis trying several values of the autocorrelation parameter, say for example .1,.2, and .3. If the results are very similar with different values of the autocorrelation parameter, then the user knows their results will still hold if there is autocorrelation present. If the results change as the autocorrelation parameter changes then the user may end up presenting the series of results, to show how the results depend on different assumptions about the autocorrelation.
3. If the user suspects negative autocorrelation, there is need to do any further analysis (see Kim, et al., 2000).

If you suspect negative autocorrelation, the uncorrelated errors model will suffice (see Kim, et al., 2000).

Note: Only the uncorrelated errors model can be used with the Test for Pairwise Differences.

Number of Observations

To keep JoinPoints from being placed too close to the end points, specify the minimum number of observations from a JoinPoint to either end of the data (including the first or last JoinPoint if it falls on an observation). This value must be at least two (2), but the default is set to three (3). So, for example if this value is set to 3 and the data are annual from 1973-2004, then:

- With 0 points between adjacent observed values (see below), the first possible JoinPoint is 1975.
- With 3 points between adjacent observed values (see below), the first possible JoinPoint is 1974.25.

In both cases there are 2 observed values before the first JoinPoint, the observations at 1973 and at 1974.

To keep JoinPoints from getting too close together, specify the minimum number of observations between two JoinPoints (including any JoinPoint that falls on an observation). This value can be set as low as two (2), but the default is set to four (4). For example, if this value is set to 4 and the data are annual from 1973-2004, then:

- With 0 points between adjacent observed values (see below), if there was a JoinPoint at 1981, then the closest possible other JoinPoints would be at either 1978 or 1984.
- With 3 points between adjacent observed values (see below), if there was a JoinPoint at 1981.75, then the closest possible other JoinPoints would be at either 1979.75 or 1983.25.

If this setting (minimum number of observations between two JoinPoints) is set to two, then the JoinPoint model is not full rank and the program will use a generalized inverse to find (non-unique) parameters that minimize the sum of squared errors.

The permutation tests are valid for any of the allowable choices for the above two settings (the minimum number of observations from a JoinPoint to either end of the data, and the minimum number of observations between two JoinPoints); however, some statistics (the standard error of the slope parameters and the associated *p*-values) cannot be calculated when there are not at least three observations on a line segment (excluding observations at the JoinPoints). See JoinPoint Output Statistical Notes. Similarly, some statistics cannot be calculated if a line segment in the JoinPoint model is an exact fit.

The program uses a grid search to find the MLE of the JoinPoints. The user may specify the coarseness of the grid by specifying the number of grid points (from zero to nine) to place between adjacent observed *x* values. If this value is zero, the grid will be the observed *x* values. If this value is one, the grid will be the observed *x* values plus the midpoints between observed *x* values, etc. Please note: If this number exceeds three or four, the execution time required by the calculation engine may be excessive.

More:
Number of JoinPoints
Statistical Notes

Model Selection Method –How JoinPoint Selects the Final Model

JoinPoint selects the optimal model using three different methods.

The first method uses the sequence of permutation tests to ensure that the approximate probability of overall Type I error is less than the specified significance level (also called the alpha level, default = .05). Assuming that the default value of the minimum number of JoinPoints is 0, "the overall Type I error" is the probability of incorrectly concluding that the underlying model has one or more JoinPoints when, in fact the true underlying model has no JoinPoints.

The second method is based on the Bayesian Information Criterion (BIC). The value of BIC is the loglikelihood value with penalizing the cost of extra parameters. The model with the minimum value of BIC is selected as the optimal model.

The third method is Modified BIC - a modification of traditional BIC proposed to improve its performance.

Here are the details for the three different methods:

- Permutation Tests
- BIC (Bayesian Information Criteria)
- Modified BIC

Permutation Tests

The program performs multiple tests to select the number of JoinPoints, using the Bonferroni correction for multiple testing. Set the overall significance level for multiple testing.

The program performs permutation tests to select the number of JoinPoints. Since fitting all N! possible permutations of the data would take too long, the program takes a Monte Carlo sample of these N! data sets, using a random number generator. Specify the size of the Monte Carlo sample of permuted data sets, or set the size equal to zero to turn off the permutation test option.

More:
How JoinPoint Conducts Permutation Testing
Permutation Test Details
JoinPoint *p*-values
Random Number Generator

How JoinPoint Conducts Permutation Testing

In the JoinPoint Regression Program, the permutation test is used repeatedly for testing between two different JoinPoint models, a simpler model with fewer JoinPoints called the null model, and a more complicated model called the alternative model. The alternative model fits better because it is more complicated.

The question for the test is: does it fit much better than would be expected by chance. To test this statistically, we calculate a ratio, SSE_N/SSE_A, where SSE_N is the sum of squared errors (SSE) from the null model and SSE_A is the SSE from the alternative model. Values of the ratio close to 1 mean that the alternative is not much better than the null model, while larger values mean that the alternative is much better.

In order to decide how much larger a ratio needs to be to be statistically significant, we use the permutation method. In this method we randomly permute (that is, shuffle) the errors (also called the residuals) from the null model and add them back onto the modeled values from the null model to create a permutation data set. Then we calculate the ratio for the permutation data set.

- If the true model was the null model we would expect that about half of the ratios calculated from the permutation data set would be less than the original one.
- If the true model were the alternative model, we would expect that after permuting the errors most of the new ratios would be less than the original ratio.

In other words the permuted data set would look less like the alternative model than the original data. So we reject the null model (or null hypothesis) if less than a certain proportion of the ratios are greater than or equal to the original ratio.

Permutation Test Details

First, the user specifies MIN as the minimum number of JoinPoints and MAX as the maximum number of JoinPoints on the Advanced tab. Then the program uses a sequence of permutation tests to select the final model. Each one of the permutation tests performs a test of the null hypothesis H_0: number of JoinPoints=k_a against the alternative H_a: number of JoinPoints=k_b. The procedure begins with k_a = MIN and k_b = MAX. If the null is rejected, then increase k_a by 1; otherwise, decrease K*b* by 1. The procedure continues until k_a=k_b and the final value of $\hat{k} = k_a = k_b$ is the selected number of JoinPoints.

Significance level of each individual test in a sequential testing procedure

Because multiple tests are performed, the significance level of each test is adjusted to control the overall type I error at specified α level (0.05). Before Version 3.0, Bonferroni adjustment was used, i.e., α1= α/(MAX - MIN) and if the individual test p-value is less than α1, the null is rejected.

The Bonferroni adjustment is conservative because the actual overall significance level is usually less than the nominal level α. Starting with Version 3.0, the new adjustment procedure controls the overall over-fitting error probabilities,

$$P(k > K_a \mid k = K_a), K_a = MIN...MAX ,$$

under α. Let k denote the number of JoinPoints and α (k_a; k_b) be the significance level of each individual test H0 : k = k_a vs. H1 : k = k_b. The new procedure set α (k_a; k_b) = α/(MAX - k_a). Notice that the individual significance level depends on the number of JoinPoints k_a under the null. Consider an example where MIN = 0 and MAX= 4. The new procedure has the following properties:

$$P(k > 0 \mid k = 0) = \alpha(0,4) + \alpha(0,3) + \alpha(0,2) + \alpha(0,1);$$
$$P(k > 1 \mid k = 1) = \alpha(1,4) + \alpha(1,3) + \alpha(1,2);$$
$$P(k > 2 \mid k = 2) = \alpha(2,4) + \alpha(2,3);$$
$$P(k > 3 \mid k = 3) = \alpha(3,4).$$

If we like to bound these over-fitting probabilities by α, then we can assign different values for each α (k_a; k_b) . That means, we can achieve a better power by setting

$$\alpha(0,4) = \alpha(0,3) = \alpha(0,2) = \alpha(0,1) = \alpha / 4;$$
$$\alpha(1,4) = \alpha(1,3) = \alpha(1,2) = \alpha / 3;$$
$$\alpha(2,4) = \alpha(2,3) = \alpha / 2;$$
$$\alpha(3,4) = \alpha.$$

More:
Significance levels in Version 2.7 and earlier

Significance level of each individual test used in Version 2.7 and earlier

JoinPoint selects the model by using an algorithm to ensure that the approximate overall Type I error is less than the specified significance level (significance level is also called the alpha level, default=0.05). Assuming that the minimum specified number of JoinPoints is the default value of 0, the "overall Type I error" is the probability of incorrectly concluding that the underlying model has one or more JoinPoints when, in fact, the true underlying model has no JoinPoints. Several different statistical tests are performed to arrive at the final model. To control the overall Type I error for the entire group of tests, each test must be conducted at a significance level less than the specified (overall) significance level.

Here are the details:
Suppose that you specify MIN as the minimum number of JoinPoints and MAX as the maximum number of JoinPoints on the **JoinPoint** tab, and you specify the overall significance level as 0.05. Then the program performs the following tests: It starts by testing the null hypothesis of MIN JoinPoints against the alternative hypothesis of MAX JoinPoints. If it rejects the null (that is if the *p*-value is less than 0.05/(MAX-MIN)) then it tests MIN+1 JoinPoint against MAX JoinPoints, while if it fails to reject then it tests MIN JoinPoints against MAX-1. It proceeds in a similar manner, where it increases the number of JoinPoints under the null hypothesis by one if the null hypothesis is rejected, or decreases the number of JoinPoints under the alternative by one otherwise, until it completes testing the null hypothesis of k JoinPoints against the alternative of k+1 JoinPoints for some MIN ≤ k < MAX. The final model uses k+1 JoinPoints if the final null hypothesis is rejected, and k otherwise. Each of these permutation test are carried out with a significance level of 0.05/(MAX-MIN), to ensure that the approximate overall probability of concluding MIN+1 or more JoinPoints when the true model has MIN JoinPoints is less than 0.05.

JoinPoint p-values

The JoinPoint Regression Program performs a series of hypothesis tests that test the null hypothesis of k_a JoinPoints against the alternative hypothesis of k_b JoinPoints, where k_a and k_b change for each hypothesis test. Each *p*-value corresponds to this type of test. The *p*-value is an estimate of the probability, under the assumption that there are only k_a JoinPoints, of observing data that look more like a JoinPoint model with greater than k_b JoinPoints than the data that we have in fact observed.

As the permutation test is a randomization test, which depends on the random number generator. For greater consistency in the permutation test *p*-values obtained if one were to change the seed for each run, we suggest running the program for at least 4499 permutations. For this reason, the default number of permutations is now 4499 in the current version of the JoinPoint Regression Program. Choice of the number of permutations selected by the user is a trade-off between computer time and consistency of the *p*-values obtained.

Random Number Generation

The program performs permutation tests to select the number of JoinPoints. Since fitting all N! possible permutations of the data would take too long, the program takes a Monte Carlo sample of these N! data sets, using a random number generator to calculate *p*-values for a series of permutation tests.

Here we discuss the implications of the choice of the number of permutation data sets, say N. The program runs faster with smaller values of N, but it gives better power with larger values of N. In addition, a larger N reduces the probability that another analysis of the same data might get a different answer when run with different random number generator seeds.

Computer programs can produce pseudo-random numbers through algorithms that mimic randomness, which we use to shuffle or permute the errors. The algorithms use a seed or seeds to start the algorithm. These seeds can be used to produce repeatable pseudo-random numbers. The problem of two analyses obtaining different answers from the same data is addressed by this program by specifying default random number generator seeds. Thus, as long as no parameters are changed (including the random number generator seed and N), repeats of the analyses will produce the same results. Otherwise, two runs of the same analysis using different seeds could get different answers.

To get an idea how results would change for someone using different random number generator seeds, we list some confidence intervals for *p*-values below. For example, with N=999 Monte Carlo samples if you obtained a *p*-value of .04 from the program there is an approximately 99% chance that another researcher repeating the analysis with N very large (i.e., an ideal situation with N -> infinity) would obtain a *p*-value between .025 and .0577.

N=99

lower 99% ci	estimate	upper 99% ci
0.0000	0.01	0.0521
0.0034	0.04	0.1065
0.0069	0.05	0.1218
0.0111	0.06	0.1364
0.0325	0.10	0.1910
0.2702	0.40	0.5281

N=999

lower 99% ci	estimate	upper 99% ci
0.0031	0.01	0.0199
0.0250	0.04	0.0577
0.0331	0.05	0.0694
0.0415	0.06	0.0810
0.0762	0.10	0.1259
0.3595	0.40	0.4402

N=9999

lower 99% ci	**estimate**	**upper 99% ci**
0.0075	0.01	0.0127
0.0350	0.04	0.0452
0.0445	0.05	0.0558
0.0540	0.06	0.0663
0.0923	0.10	0.1079
0.3873	0.40	0.4127

N=99999

lower 99% ci	**estimate**	**upper 99% ci**
0.0092	0.01	0.0108
0.0384	0.04	0.0416
0.0482	0.05	0.0518
0.0581	0.06	0.0620
0.0976	0.10	0.1025
0.3960	0.40	0.4040

Typically, you should allow the computer to use the default seeds for the random number generator. By using these default seeds, one can duplicate results from a previous JoinPoint session even though the program uses Monte Carlo sampling. Although not recommended in general, the JoinPoint Regression Program allows one to change the default seed by selecting **Session > Preferences** from the **File** menu.

Early Stopping Options

Since Hudson's Method is computationally intensive, Early Stopping Options have been added so that not all permutations need to be analyzed. The options are:

- Fixed – all permuted data sets are analyzed, with the default being 4499.
- B-Value – the maximum number of permuted data sets to be analyzed is determined by the significance level specified, using a less conservative approach. See Fay et al. for details (Fay, MP, Kim, H-J, and Hachey, M. (2007) "On using Truncated Sequential Probability Ratio Test Boundaries for Monte Carlo Implementation of Hypothesis Tests" (to appear in Journal of Computational and Graphical Statistics).
- Curtailed – the maximum number of permuted data sets to be analyzed is determined by the significance level specified, using a more conservative approach.

This option is not currently available for the Grid Search Method but may be added in a future version of the software.

More:
References

BIC (Bayesian Information Criterion) Details

The equation for computing the BIC for a k-JoinPoint model is:

BIC(k) = ln{SSE(k)/#Obs(k)} + {#Parm(k) /#Obs} * ln(#Obs),

where SSE is the sum of squared errors of the k-JoinPoint regression model, #Parm(k)=2*(k+1) is the number of parameters of the k-JoinPoint model and #Obs is the number of observations.

The k-JoinPoint model with the minimum value of BIC(k) is selected as the final model.

Modified BIC

Zhang and Siegmund (2007, Biometrics) discussed that in the context of change-point problems, the traditional BIC does not satisfy the technical assumptions of Schwarz (1978, Annals of Statistics) and proposed a modification to improve its performance. The MBIC in JoinPoint regression is derived as an asymptotic approximation of the Bayes factor and is of the form:

$$\text{MBIC(k)} = \text{BIC(k)} + \frac{\ln|X_k'(\hat{\tau})X_k(\hat{\tau})|}{n} - \frac{2}{n}\, ln\Gamma\left(\frac{n-k-3}{2}\right) - \frac{k+3}{n}\ln\big(SSE(k)\big),$$

where n is the number of observations, Γ(z) is the gamma function:

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\, dt$$

and

$$X_k(\hat{\tau}) = \begin{pmatrix} 1 & x_1 & (x_1-\hat{\tau}_1)^+ & \dots & (x_1-\hat{\tau}_k)^+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n-\hat{\tau}_1)^+ & \dots & (x_n-\hat{\tau}_k)^+ \end{pmatrix}$$

for a k-JoinPoint model with the values of the independent variable $(\chi_1, \dots, \chi_\eta)$ and the JoinPoints estimated as $(\hat{\tau}_1, \dots, \hat{\tau}_k)$ Note that a+ = max (a,0).

Session Preferences

Access the Session Preferences window by selecting **Session Preferences** from the Session menu (only available when a session is open). Make changes to the fields as necessary. When you are finished, click **OK**.

Alpha Levels

The alpha levels for the JoinPoint Locations, APCs, and AAPCs are set at 0.05 by default. These values can be changed here.

Random Number Generation

Typically, you should allow the computer to use the default seeds for the random number generator. By using these default seeds, one can duplicate results from a previous JoinPoint session even though the program uses Monte Carlo sampling. Although not recommended in general, the JoinPoint Regression Program allows one to change a default seed here..

Executing the JoinPoint Program

Once you have set all controls in the JoinPoint Regression Program, execute the program by clicking ⚡ on the toolbar or clicking Run on the menu bar.

A progress meter will be shown on the screen while the JoinPoint calculation engine processes the data and generates the output. Click Cancel if you would like to halt the calculations before they are complete.

Once the processing is complete, an output window will automatically be displayed. You will need to save this output if you would like to access it again at a later time.

More:
JoinPoint Output
JoinPoint Output – Statistical Notes

JoinPoint Output

Click on the various tabs (**Graph**, **Data**, **Model Estimates**, or **Model Selection**) to view graphs, view the graph data, or view the report. You may also access both the graph data and the report information by exporting these for use by other software packages (Excel, Harvard Graphics, SAS, etc.).

The JoinPoint calculation engine displays the output window but doesn't automatically save or create a permanent output file. You can save the output as a single binary file and/or you can export the graphs as bitmap files and the graph data and the report information as text files. JPO (JoinPoint Output) is the extension JoinPoint uses for the output files.

Select **File > Print…** to send a report of the ouput to a printer or a PDF file. The **Printing Options** window allow you to select which information to include in the report.

More:
Graph
Data
Model Estimates
Trends
Model Selection
Output Properties
Output Options
Exporting Results
Statistical Notes

Output Window - Graph Tab

The JoinPoint Regression Program saves the graph coordinates for the observed and predicted data in the output file and uses this information to display the graphs and data tables. You may export this data for use in a commercial graphics package such as Harvard Graphics or Excel.

To manipulate the titles, labels, colors, and/or axis scaling of the graph:

- Select **Output > Options..**. from the top of the main JoinPoint window.
- Click on the **Graph** tab.
- There are sub-tabs for each display option (**X-Axis**, **Y-Axis**, **Appearance**, **Titles/Labels**, **Scale**). Make your changes and click **OK** or **Apply**. The graph will be updated to reflect your selections.

More:
Data
Model Estimates
Model Selection

Output Window - Data Tab

The JoinPoint Regression Program writes the graph coordinates for the observed and predicted data to the Output File. The JoinPoint Regression Program uses this file to display the graphs and data tables. You may also import this file into a commercial graphics package such as Harvard Graphics or Excel.

To view your output in a data table:

- Click on the **Data** tab.
- You can change the number of decimal places displayed with **Output > Options** and, also, whether or not the APC or Slope is displayed.

More:
Graph
Model Estimates
Model Selection

Output Window – Model Estimates Tab

The Model Estimates Tab of the JoinPoint Output Window displays the specifications for each of the models run in the analysis. The Cohort control at the top of the Output Window selects which By-Group to display, and the Model control selects which model (i.e. the 1 JoinPoint model or the 3 JoinPoint model) to display. The best, selected model is first displayed.

To control which statistics are displayed, or to change the number of decimal places displayed:

- Select **Output > Options...** from the top of the main JoinPoint window.
- Click on the **Model Estimates** tab.
- Change the number of decimal places for each type of statistic. Click or unclick the box by each type of statistic you would like displayed or hidden. Click **OK** or **Apply** after you have made your changes.

You may also export this data for use in another software package. Note, only what is displayed in the Output Window will be exported. For example, if only 1 decimal place is chosen, only that many decimal places will be in the exported file. Or, if the "Regression Coefficients" box is unchecked, those statistics will not be included in the exported file.

More:
Data
Graph
Trends
Model Selection

Output Window – Trends Tab

The Trends Tab of the JoinPoint Output Window displays the estimated JoinPoints, APCs, and AAPCs for each of the models run in the analysis. Note: this window is only displayed if the log-linear model {ln(y) = xb}was selected. The Cohort control at the top of the Output Window selects which By-Group to display, and the Model control selects which model (i.e. the 1 JoinPoint model or the 3 JoinPoint model) to display. The best, selected model is first displayed.

To control which statistics are displayed, or to change the number of decimal places displayed:

- Select **Output > Options...** from the top of the main JoinPoint window.
- Click on the **Trends** tab.
- Change the number of decimal places for each type of statistic. Click or unclick the box by each type of statistic you would like displayed or hidden. Click **OK** or **Apply** after you have made your changes.

To control which AAPCs are displayed:

- Select **Output > Specify AAPCs...** from the top of the main JoinPoint window (or **Specify AAPCs...** on the **Trends** tab).
- Then, follow the instructions identified in the help for Specify AAPCs.

You may also export this data for use in another software package. The APCs are included in the export of the Model Estimates. To export the AAPCs, select **Output > Export As Text > AAPCs...**, choose a filename and any desired options, and click **OK**.

More:
Model Estimates
Specify AAPCs
Average Annual Percent Change (AAPC)

Specify AAPCs

Since AAPCs are computed after the calculation engine has run, based on the APCs, they can be modified in the output window without rerunning the session. To specify which AAPCs to display, select **Output>Specify AAP Ranges...** from the menu and check the box for the entire range of data or the last *N* observations or the first and last endpoints of the range for which the AAPC should be computed. Note: the endpoints must occur on actual observations.

More:
Average Annual Percent Change (AAPC)

Output Properties

Select **Output > Properties** from the top of the main window to see information on the settings used to create the active output file. This also displays how long it took to run the session.

If you want to re-run the models with different settings, a convenient way is to retrieve the session from the output file.

More:
Retrieve Session from an Output File

Retrieving a Session used to create JoinPoint Output

The JoinPoint session file is saved as part of the output, so you can select **Output > Retrieve Session** from the top of the main window to retrieve the session that was used to create the output, even if that session file wasn't specifically saved. This is a convenient way to see exactly what settings were used to create a particular output file.

Options for displaying JoinPoint Output

Select the **Output > Options** from the **File** menu to customize the output that is being displayed. Options include:

- modifying the titles, labels, colors, and/or axis scaling of the graphs.
- the number of decimal places to display for various statistics.
- which statistics, or sections, to include (or hide).

Note: The options you select will be saved with that particular output file (if you save the file after changing the options). If you would like your preferences to always be the default for any new JoinPoint runs, click on **Set as Default** after you have set the options the way you desire.

Exporting Results from JoinPoint Output

You may access both the graph data and the report information by exporting these as text files for use with other software packages (Excel, Harvard Graphics, SAS, etc.).

To export a graph that is being displayed as an image file (picture), select **Output > Export Graph...** from the **File** menu and choose the name and location of the bitmap file to be saved.

To export information (data) from a JoinPoint run, first select **Output > Export Text** from the **File** menu and then, depending on the data you desire, select:

- **Data...** – Choose this option to export the observed and predicted values for each observation.
- **Model Estimates** – Choose this option to export the model estimates and related information. Select **Table** to put the data in rows that can be used for Excel.
- **Trends (APC)...** – Choose this option to export the APCs.
- **Trends (AAPC)...** – Choose this option to export the AAPCs.
- **Permutation Tests... or BIC...** – Choose one of these options to export information about the model selection process.

More:
Tips for Working with Excel
Printing Options (JoinPoint Output)

Tips for working with Excel

To import information from a JoinPoint run into Excel, first, export the data using **Output > Export as Text** and then, simply use the **File** menu in Excel and Open the exported file. Step through the Excel "Text Import Wizard". Specify that it is a delimited file. Then specify the delimiter you chose in the export process.

Statistical Notes

Overview

Let K_{min} and K_{max} be the minimum and maximum for the number of JoinPoints, respectively. First the program goes through each of the k-JoinPoint models, $K_{min} \leq k \leq K_{max}$. For each of the models, the program chooses the regression parameters with the smallest sum of squared error (SSE, or smallest weighted SSE). Statistics related to each of the k-JoinPoint models are discussed in the following sections . The sequential permutation test procedure to choose the best JoinPoint model is discussed in detail in Kim et al. (2000), and only briefly described here.

More:
Statistics Related to the k-JoinPoint Model
Significance Level of each Individual Test in a Sequential Testing Procedure
References

Statistics Related to the k-JoinPoint Model

Several statistics related to the k-JoinPoint model are described here. All statistics come directly from Lerman (1980).

More:
Notation
Parameterizations
SSE and MSE
Estimated JoinPoints
Estimated Regression Coefficients (Beta)
Estimated APC
AAPC

Notation

- n is the total number of data points.
- k is the number of JoinPoints in the model.
- p is the total number of parameters in the model, including the JoinPoint parameters. For our models p=2k+2.
- $Q_k$is the sum of squared errors (SSE) from the model that minimizes SSE with k JoinPoints. If a weighted analysis is done (anything but heteroscedastic errors option= constant variance) then Q_k represents the minimum weighted SSE.
- $Q_{x,j,k}$ is the (weighted) SSE from the model that minimizes (weighted) SSE with k JoinPoints and with the jth JoinPoint occurring at x.
- $\mathbf{F^{-1}_{a,b}(p)}$ is the pth quantile of the F distribution with a and b degrees of freedom.

Parameterizations

The program outputs parameters from two different parameterizations: the "general changepoint" parameterization (GCP), and the "standard" parameterization (SP) of Kim *et al.* (2000).

The standard parameterization is (see Kim, et al 2000, equation 1),

E[y|x] = $\beta_0 + \beta_1$ x + δ_1 (x- τ_1)+ + ... + δk (x - $\tau_k)^+$ (1)

where $(a)^+ = a$ if $a > 0$ and 0 otherwise.

The general changepoint parameterization is,

$$E[y|x] = \sum_{j=1}^{k+1} (\beta_{0j} + \beta_{1j}\, x)\, I\, (\tau_{j-1} < x \leq \tau_j) \qquad (2)$$

where $\tau_0 = min(x)$ and $\tau_{k+1} = max(x)$, and under the constraint that $E[y|x]$ is continuous at τ_j.

For the relationship between the parameterizations see Table 1 and Appendix A.

Table 1: Parameter transformations for different models

Output Label	**Standard**	**General Changepoint**
Intercept 1	β_{01}	β_0
Intercept j, j > 1		$\beta_0 - \Sigma_{l=1}^{j-1}$ rldl
Slope 1	β_{11}	β_1
Slope j, j > 1		$\beta_1 + \Sigma_{l=1}^{j-1} \delta_l$
Slope j - Slope (j-1), j ≥ 2	$\beta_{1,j} - \beta_{1,j-1}$	

More:
Appendix A

Appendix A

Relationship Between Standard and General Changepoint Parameterizations

Rewrite standard,

$$\begin{aligned} E[y|x] &= \beta_0 + \beta_1 x + \delta_1 (x - \tau_j)^+ + \ldots + \delta_k (x - \tau_k)^+ \\ &= \beta_0 + \beta_1 x + \sum_{j=1}^{k} \delta_j (x - \tau_j) I(x > \tau_j) \\ &= \beta_0 + \beta_1 x + \sum_{j=1}^{k} \delta_j x I(x > \tau_j - \sum_{j=1}^{k} \delta_j \tau_j I(x > \tau_j) \end{aligned}$$

Rewrite GCP,

$$\begin{aligned} E[y|x] &= \sum_{j=1}^{k+1} (\beta_{0j} + \beta_{1j}\, x) \{ I(x > \tau_{j-1}) - I(x > \tau_j) \} \\ &= \sum_{j=1}^{k+1} (\beta_{0j} + \beta_{1j}\, x)\, I(x > \tau_{j-1}) - \sum_{j=1}^{k+1} (\beta_{0j} + \beta_{1j}\, x)\, I(x > \tau_j) \\ &= \sum_{j=0}^{k} (\beta_{0,j+1} + \beta_{1,j+1} x)\, I(x > \tau_j) - \sum_{j=1}^{k} (\beta_{0j} + \beta_{1j} x) I(x > \tau_j) \\ &= \beta_{01} + \beta_{11} x + \sum_{j=1}^{k} I(x > \tau_j) \{ (\beta_{0,j+1} - \beta_{0j}) + (\beta_{1,j+1} - \beta_{1j}\, x) \} \end{aligned}$$

So that

Standard	**General Changepoint Parameterizations**
β_0	$= \beta_{01}$
β_1	$= \beta_{11}$
δ_j	$= \beta_{1,j+1}$ - $\beta_{1j}{}^{x}$
r_j	$= \dfrac{\boldsymbol{\beta_{0j+1} - \beta_{0j}}}{\boldsymbol{\beta_{1j+1} - \beta_{1j}}}$

SSE and MSE

Sum of squared errors (SSE) is actually the weighted sum of squared errors if the heteroscedastic errors option is not equal to constant variance. The mean squared error (MSE) is the SSE divided by the degrees of freedom for the errors for the constrained model, which is n-2(k+1).

The minimum SSE for a k-JoinPoint model is calculated using Lerman's grid-search method (1980) based on Kim et al's standard parametrization (Equation 1). The corresponding values for (τ_1,..., τ_k) and (β_0, β_1, δ_1,..., δ_k) are the estimates of JoinPoints and regression coefficients, respectively.

Degrees of Freedom

When the Grid Search Method is used, the degree of freedom for the estimated regression coefficient is d=n*-2(k+1), where k is the number of JoinPoints and n* is the effective number of data points after deleting offending observations, data points that are on JoinPoints.

When Hudson's Method is used for a continuous fitting, the degree of freedom is adjusted in a continuous manner instead of subtracting the offending observations. The idea is to delete, with higher probabilities, the observations whose x-values are closer to JoinPoints, while no more than one observation is deleted around each JoinPoint. See Kim et al. (Kim, H.-J., Yu, B. and Feuer, E.J. (2007) "Inference in segmented line regression: A simulation study", to appear in Journal of Statistical Computation and Simulation) for details.

More:
References

Estimated JoinPoints

If *k* > 0 then the output lists the estimated JoinPoints. The associated confidence intervals (CI) come from Lerman (1980) equation 6 using $C_2\,\alpha$. Explicitly, the100(1 α)% confidence interval for the *j*th of *k* JoinPoints includes all values of *x* from the grid such that $Q_{x,j,k} \leq C_2\,\alpha$, where

$$C_2^{\alpha} = Q_k\left(1 + \frac{k}{n-p}\right)F_{k,n-p}^{-1}(1-\alpha)$$

Estimated Regression Coefficients (Beta)

The output is a combination of the two parameterizations (see Table 1). The estimates of (β_0, β_1, δ_1,..., δ_k) come from the grid-search method. The estimates of (β_{00}, β_{01},..., $\beta_{0,k+1}$, $\beta_{1,k+1}$) are calculated based on Table 1.

However, the standard errors of the regression coefficients are estimated under the GP model (Equation) without continuity constraints. Following Lerman's implementation, (Lerman; 3rd paragraph, page 79, 1980; Feder, Section 4, page 69, 1975), the data points that are on the JoinPoints are deleted. Then conditioned on the partition implied by the estimated JoinPoints (τ_1,..., τ_k), the standard errors of (β_{00}, β_{01},..., $\beta_{0,k+1}$, $\beta_{1,k+1}$) are calculated using unconstrained least square for each segment. If there are segments with zero or one observation (not including the JoinPoints at both ends), then a generalized inverse is used to calculated the covariance matrix. The standard error of the difference in slopes, δ_j, is the square root of the sum of the squared standard errors (variance) for the two consecutive slopes β_{1j} and $\beta_{1,j+1}$.

The Z statistic is the parameter estimate divided by the standard error. The Z-statistic has a t distribution with d degrees of freedom. Let n_J be the number of data points that are on JoinPoints. The effective number of data points is $n^*=n-n_J$. Let k_0 and k_1 be the number of segments with zero or one observation. The effective number of parameters (rank of the design matrix) is $p^*=2(k+1)-2k_0-k_1$ and the degrees of freedom $d=n^*-p^*$. For the default option where the minimum number of data points between two JoinPoints (excluding any JoinPoint that falls on an observation) is two, $k_0=k_1=0$ and $d=n^*-2(k+1)$. For testing H_0: β = 0 the p-value is calculated as Prob > $|Z|=2\{1-t_d(|Z|)\}$, where t_d is a t distribution with d degrees of freedom.

Estimated APC

When the model option on the Specifications tab is *ln*(*y*)=*xb*, then the output calculates the estimated annual percentage rate change (APC). For any segment with slope β the APC is 100{ *exp*(β) -1 }. The 100(1- α)% confidence limits are:

Lower = 100{ exp(β - s*t_d^{-1}(1 α/2)) -1 }

Upper = 100{ exp(β + s*t_d^{-1}(1 α/2)) -1 }

Upper = 100{ exp(β /2)) -1 }

where d is the degree of freedom and s is the standard error for the slope listed in the output (i.e., from the unconstrained linear models), and $t_d^{-1}(q)$ is the qth quantile of a t distribution with d degrees of freedom.

Average Annual Percent Change

Average Annual Percent Change (AAPC) is a summary measure of the trend over a pre-specified fixed interval. It allows us to use a single number to describe the average APCs over a period of multiple years. It is valid even if the JoinPoint model indicates that there were changes in trends during those years. It is computed as a weighted average of the APC's from the JoinPoint model, with the weights equal to the length of the APC interval. For more details, see: http://surveillance.cancer.gov/JoinPoint/aapc.html.

More:
Specify AAPCs

AAPC Comparison

When the test for pairwise differences is selected, and the two groups are not parallel or coincident, the AAPCs for the two groups are tested to determine if they are statistically different. For details, see http://surveillance.cancer.gov/JoinPoint/aapc.html.

Significance Level of each Individual Test in a Sequential Testing Procedure

The JoinPoint program uses a sequence of "permutation" tests to select the final model. Each one tests the null hypothesis H_0: $k=k_a$ against the alternative hypothesis H_1: $k=k_b$. The procedure begins with $k_a=K_{min}$ and $k_b=K_{max}$. If the null is rejected, then increase k_a by 1; otherwise, decrease k_b by 1. Let $\hat{k} = k_a = k_b$ be the final selected number of JoinPoints.

Because multiple tests are performed, Bonferroni adjustment is used to ensure that the approximate overall type I error is less than the specified significance level (significance level is also called the α-level, default α=.05). Each of these permutation test are carried out a significance level of $\alpha_1= \alpha/(K_{max}-K_{min})$, i.e., if the p-value < α_1, then it rejects the null.

The Bonferroni adjustment is conservative because the actual overall significance level is usually less than the nominal level α. The new adjustment procedure controls the overall over-fitting probabilities P([^k] > k_a|$k=k_a$), $k_a=K_{min},...,K_{max}$ under α. Let α (k_a,k_b) be the significance level of each individual test H_0: $k=k_a$ vs. H_1: $k=k_b$. The new procedure set α (k_a,k_b)= α /(K_{max}-k_a). Notice that the individual significance level depends on the number of JoinPoints k_a under the null. Consider an example where K_{min}= 0 and K_{max}= 4. The new procedure has the following properties:

$$P(k > 0 \mid k = 0) = \alpha(0,4) + \alpha(0,3) + \alpha(0,2) + \alpha(0,1);$$
$$P(k > 1 \mid k = 1) = \alpha(1,4) + \alpha(1,3) + \alpha(1,2);$$
$$P(k > 2 \mid k = 2) = \alpha(2,4) + \alpha(2,3);$$
$$P(k > 3 \mid k = 3) = \alpha(3,4).$$

If we like to bound these over-fitting probabilities by α, then we can assign different values for each α (k_a,k_b). That means, we can achieve a better power by setting

$$\alpha(0,4) = \alpha(0,3) = \alpha(0,2) = \alpha(0,1) = \alpha / 4;$$
$$\alpha(1,4) = \alpha(1,3) = \alpha(1,2) = \alpha / 3;$$
$$\alpha(2,4) = \alpha(2,3) = \alpha / 2;$$
$$\alpha(3,4) = \alpha.$$

References

Basic Method

Kim, H-J, Fay, M.P., Feuer, E.J., and Midthune, D.N. (2000) "Permutation Tests for JoinPoint Regression with Applications to Cancer Rates", Statistics in Medicine 19, 335-351. (correction: 2001;20:655). Correction to Table 1(a) of Kim, et al. is provided as a PDF at http://surveillance.cancer.gov/documents/JoinPoint/table1.pdf.

> This is a paper where JoinPoint regression is applied to describe cancer rates and the permutation test is proposed to determine the number of significant JoinPoints. The grid search of Lerman (1980) was used to fit the segmented regression function and the p-value of each permutation test is estimated using Monte Carlo methods, and the overall asymptotic significance level is maintained through a Bonferroni adjustment.

Background Papers for Basic Method

Feder, P.I. (1975) "On Asymptotic Distribution Theory in Segmented Regression Problems: Identified Case", Annals of Statistics 3, 49-83.

> Feder studied asymptotic properties of the least squares estimators in multi-segment regression and proved that under some technical conditions on the independent variable, the least squares estimators are consistent and asymptotically normal.

Lerman, P.M. (1980) "Fitting Segmented Regression Models by Grid Search" Applied Statistics, 29: 77-84.

> Lerman proposed the grid search method to fit segmented line regression where the JoinPoint estimates occur at discrete grid points, and studied asymptotic inference using asymptotic normality proved by Feder (1975).

Hudson, D. (1966) "Fitting segmented curves whose join points have to be estimated", Journal of the American Statistical Association 61, 1097-1129.

> Hudson proposed a procedure to fit a segmented regression curve whose JoinPoints can be estimated anywhere in the data range. In the first stage, the model is fit to every feasible partition of the data without imposing continuity across the change-points. We then examine the locations of the intersection points of the least square lines. When the intersection point of the two adjacent least squares lines is not between the two data points which separate the segments, we make an adjustment in the fit by mathematically imposing the continuity constraints at one of the two boundary data points. The final estimates are obtained by searching for the global minimum of residual sum of squares.

Fitting JoinPoints in Continuous Time

Yu, B., Barrett, M., Kim, H-J, and Feuer, E.J. (2007) "Estimating JoinPoints in Continuous Time Scale for Multiple Change-Point Models", Computational Statistics and Data Analysis 51, 2420-2427.

> In this paper, we extend the Hudson's continuous fitting method to a multiple JoinPoint model and discuss some practical issues in the implementation. We also compare computational efficiencies of the Lerman's grid search and the Hudson's continuous fitting.

Early Stopping Rules for the Permutation Test

Fay, M.P., Kim, H-J, and Hachey, M. (2007) "On using Truncated Sequential Probability Ratio Test Boundaries for Monte Carlo Implementation of Hypothesis Tests", Journal of Computational and Graphical Statistics 16 (4), 946-967.

> JoinPoint selects a final model conducting a series of permutation tests and we can save computation time by using sequential stopping boundaries. This paper proposes a truncated sequential probability ratio test boundary to stop resampling when the early replications indicate a large enough or small enough p-value, and studies its properties.

Comparing Two JoinPoint Regression Lines

Kim, H-J, Fay, M.P., Yu, Binbing, Barrett, M.J., and Feuer, E.J. (2004) "Comparability of segmented line regression models", Biometrics 60, 1005-1014.

> We proposed a procedure to compare two segmented line regression functions, especially to test (i) whether two segmented line regression functions are identical or (ii) whether the two mean functions are parallel allowing different intercepts. A general form of the test statistic is described and then the permutation procedure is proposed to estimate the p-value of the comparability test.

AAPC

Clegg, L.X., Hankey, B.F., Tiwari, R., Feuer, E.J., Edwards, B.K. (2009) "Estimating average annual percent change in trend analysis". Statistics in Medicine 28(29): 3670-8.

Studies on the performance of JoinPoint

Kim, H-J, Yu, B., and Feuer, E.J. (2008) "Inference in segmented line regression: A simulation study", Journal of Statistical Computation and Simulation 78:11, 1087-1103.

> Via simulations, this paper empirically examines small sample behavior of asymptotic confidence intervals and tests, based on Feder (1975)'s asymptotic normality of least squares estimators in JoinPoint regression, studies how the two fitting methods, the grid search and the Hudson's continuous fitting algorithm affect these inferential procedures and also assesses the robustness of the asymptotic inferential procedures.

Kim, H-J, Yu, B., and Feuer, E.J. (2009) "Selecting the number of change-points in segmented line regression", Statistica Sinica 1:19(2):597-609.

> In this paper, we show that under some conditions, the number of JoinPoints selected by the permutation procedure of JoinPoint is consistent. Via simulations, the permutation procedure is compared with some information-based criteria such as Bayesian Information Criterion (BIC).

What Version am I Running?

Version 3.5 is the current version of the JoinPoint Regression Program. To verify the version number of your executable select **About** on the **Help** menu.

All changes made to the program are documented for each version release.

Please note that some versions include changes to the calculation engine. Results from these versions will differ from previous versions of the program. If you would like a previous version of the program in order to exactly replicate prior results, please request a copy of the version you would like through technical support.

More:

Version 3.5.1
Version 3.5
Version 3.4.3
Version 3.4.2
Version 3.4
Version 3.3
Version 3.2
Version 3.1
Version 3.0
Version 2.7
Version 2.6
Version 2.5.2
Version 2.5.1
Version 2.5
Version 2.4
Version 2.3
Version 2.2
Version 2.1 (beta)
Version 2.0
Version 1.1b
Version 1.1a
Version 1.0b
Version 1.0a

Version 3.5.1

Changes for version 3.5.1:

- Added alpha level to footnote on Data tab.
- Fixed spacing/alignment of Permutation Test Options.
- Removed “(see help for details)” on Permutation Tests tab.
- Reduced blank space between Two-Group tables on Permutation Tests tab.
- Changed “Help>JoinPoint on the Web>SRAB” to “>Surveillance Research Program.”
- Fixed Regional settings decimal separator difference.
- Fixed error with adding a by-variable when one doesn’t really exist.
- Fixed error when opening V3.4.3 or earlier JPO files in V3.5.
- Fixed error with comparison line segment colors on the graph.

Version 3.5

Version 3.5 was released in April 2011. The Changes included:

The Autocorrelated Errors Options were re-enabled, including a new option which estimates the autocorrelation parameters based on the data. Note: only the uncorrelated errors model can be used with the test for pairwise differences.

The confidence intervals for the AAPC were modified to follow the t-distribution, and be identical to the CIs for the APC, when the range for the AAPC falls entirely within a single JoinPoint segment.

A statistical test was added for the comparison of AAPCs between 2 groups when the pairwise differences option is selected.

Under **Session Preferences**, the significance levels can now be changed for the JoinPoint locations, APCs, and AAPCs. The defaults are 0.95.

The printing and exporting capabilities were enhanced. The items to be included in a report are selected, previewed, and then sent to a printer or a PDF file. Output can also be exported to a number of other formats like Excel, HTML, and CSV.

In versions prior to 3.5, the default for the maximum number of JoinPoints was a fixed value, most recently at 4. In this version, the maximum number of JoinPoints is determined based on the number of data points (the algorithm is described in the help). This is just a default value and can be changed by the user if desired.

Modified BIC was added as a 3rd option for the Model Selection Method. The wording for the Heteroscedastic Errors Options changed from “Poisson Model...” to “Poisson Variance…” to help clarify that these options do not use Poisson regression but instead use Weighted Least Squares regression with the assumption that the random errors are Poisson.

Version 3.4.3

Version 3.4.3 was released in April 2010. The changes Included:

An error was corrected in the P-values for the comparison test when the two groups being compared were exactly identical. This error did not affect any comparisons in which the two groups varied, even slightly.

An error was corrected that occurred when reading in SEER*Stat export files with the missing value set to zero.

An error was corrected that could occur when exporting data.

Version 3.4.2

Version 3.4.2 was released October 2009. the changes included:

A correction was made to an error in version 3.4 released in September 2009. The error was in the model fitting calculations that caused the modeled values to be off by a small factor. Although this error is unlikely to change the final model that was selected, the intermediate results and model fits will be slightly different with the corrected version. It is highly recommended that any analyses run in September 2009 with version 3.4 be re-run using 3.4.2 or a later version.

An error was corrected with the AAPC export when there are no by-variables.

Version 3.4

Version 3.4 was released in September, 2009. The changes made to the program included:

The two-group comparison was added. See Test for Pairwise Differences for details.

The maximum number of JoinPoints was increased to 9 when the Grid Search method is selected. See Number of JoinPoints for details.

The default maximum number of JoinPoints was increased to 4, to match what was used in JoinPoint runs for the SEER Cancer Statistics Review (CSR).

Cohorts with zeros in the dependent variable can now be graphed, but they will still not be modeled when the log-linear model is selected.

New shortcut icons were added to interface, for help and output options.

Warning dialog boxes are displayed when specifications indicate that the run may take a long time. These warnings can be disabled.

The autocorrelated errors option was disabled due to concerns about its algorithm. This will likely be re-enabled in future versions.

Version 3.3

Version 3.3 was released in April 2008. The changes made to the program included:

The capacity to compute Average Annual Percentage Change (AAPC) was added. See AAPC for details.

A Trends tab was added to the Output Window to display APCs and AAPCs. In previous versions, the APCs were displayed on the Model Estimates tab.

The ability to shift the Independent Variable by a fixed value was added.

The screens for changing Display Options were updated.

Version 3.2

Version 3.2 was released in January 2008. The changes made to the program are described below:

Hudson's Method was added as an alternative to the Grid Search Method. Details about this are described in Hudson's Method.

The Early Stopping Method was added as an option for the Permutation Test when Hudson's Method is selected. See Early Stopping Options for details.

The language was modified for the "Number of observations" section to accommodate Hudson's Method, but the calculations haven't changed. The JoinPoints are included in the count when they previously were excluded. So the defaults changed from 2/2/0 to 3/4/0, but the results will be the same for an annual grid search. These settings may need to be adjusted if trying to duplicate a model using a fine grid search. If you open a session from version 3.0 or earlier, these settings will automatically be updated to reflect the new language.

Data items in the Output Window can now be selected as spreadsheet-style cells and copied to other software packages like Excel.

The Model Estimates Tab of the Output Window separates the Estimated Regression Coefficients (Beta) into the two different parameterizations that are used: Standard and General.

The program can now correctly interpret decimal numbers entered in the interface when the Windows "Regional Options" is set to use a comma for decimal numbers (as in common in Europe) rather than a period (which is common in the United States).

An error in the interface for entering the autocorrelation parameter has been corrected.

Version 3.1

Version 3.1 was released in August 2007. The changes made to the program are described below.

Hudson's Method was added as an alternative to the Grid Search Method. Details about this are described in Hudson's Method.

The Early Stopping Method was added as an option for the Permutation Test when Hudson's Method is selected.

Data items in the Output Window can now be selected as cells like in a spreadsheet and copied to other software packages like Excel.

The program can now correctly interpret decimal numbers entered in the interface when the Windows "Regional Options" is set to use a comma for decimal numbers (as in common in Europe) rather than a period (which is common in the United States).

Version 3.0

Please note that the changes implemented in this version affect the results for some jobs (see the first item below).

Version 3.0 was released in April 2005. The changes made to the program are described below.

The calculation engine was coded in C++ instead of Fortran, and as a result, the random number generator and stream changed. This change affects the series of permutations, so the final model selected could change. This means that models run with a previous version of the program are not exactly reproducible with this version.

The significance level of each individual test in a sequential testing procedure was refined. For more details, see "JoinPoint Output - Statistical Notes" in the help system.

BIC (Bayesian Information Criterion) was added as an alternative method for the final model selection.

The independent variable may contain non-integer values.

Trend runs from SEER*Stat can now be loaded into JoinPoint.

The JoinPoint session file containing the specifications for the input data and model to be run is now saved as a binary file (with a .JPS extension) instead of as a text file (called a "parameter file" in versions 2.7 and earlier, with a .PAR extension). The information in these files can be exported to text files in a variety of different formats.

The output file containing the results from a JoinPoint run is now saved as a single binary file (with a .JPO extension) instead of two separate text files (.JPR and .JPO). The information in these files can be exported to text files in a variety of different formats.

Compressed .gz files can be read as the input file, and text files can be exported in .gz format.

The graphing capabilities and options have been enhanced.

JoinPoint is now an MDI (Multiple Document Interface) application, so more than one JoinPoint run can be opened in separate windows within the main JoinPoint window.

The technical support address was changed to JoinPoint@imsweb.com.

Version 2.7

Please note that the changes implemented in this version affect the results for some jobs (see item number 2 below).

Version 2.7 was released in September 2003. The changes made to the program are described below.

The calculation engine was updated to correct a problem that could occur if one of the models, within a run of multiple by-groups, had a sum of squared errors equaling zero.

The calculation of the standard errors and p-values in the Estimated Regression Coefficients (β) section of the report has been changed:

There has been a change in the tests and confidence intervals related to slopes of the JoinPoint lines or APCs. The degrees of freedom for those tests and confidence intervals have been changed to match those proposed by Lerman (1980). In addition, the standard errors, which use the degrees of freedom estimate, change accordingly. These changes will result in slightly more liberal tests or slightly smaller confidence intervals. In other words, the modified tests are a little more likely to detect statistical significance. Note that Lerman's (1980) method is justified for large numbers of observations only, and users should interpret these tests and confidence intervals cautiously with small sample sizes.

Note also that none of these changes affect the models selection procedure, which involves the series of permutations test. So the final model selected does not change and the parameter estimates do not change.

Reference

Lerman, P. M. 'Fitting segmented regression models by grid search', Applied Statistics, 29, 77-84 (1980).

Version 2.6

Please note that the changes implemented in this version affect the results for some jobs (see item numbers 3 and 5 below).

Version 2.6 was released in March 2002. The changes made to the program are described below.

JoinPoint does not allow the independent variable to have non-integer or negative values. An error message is displayed when JoinPoint encounters this situation in the input data file. The wording of this message was revised to make it more clear. (Note: future versions of the software may allow non-integer values for the independent variable.)

Text describing two parameters on the **JoinPoints** tab were revised to include a more specific description. The new labels are:

Minimum number of observations from a JoinPoint to either end of the data – (including the end data point but excluding the JoinPoint if it falls on an observation).

Minimum number of observations between two JoinPoints – (excluding any JoinPoint that falls on an observation).

The calculation of the *p*-values in the Estimated Regression Coefficients (β) section of the report has been changed from:

$$p\text{-value} = 2\ (\ 1 - \Phi\ (|Z|)\ )$$

to:

$$p\text{-value} = 2\ (\ 1 - t\ (|Z|)\ )$$

where:

Z = Parameter estimate/Standard error

Φ () is the cumulative distribution function for the standard normal distribution

$t()$ is the cumulative distribution function for the t distribution with $n - 2k + 2$ degrees of freedom

n is the total number of data points

k is the number of JoinPoints in the model

Also, the label was changed from "*Prob* > |*Z*|" to "*Prob* > |*t*|".

The "Standard Error", "*Z*", and "*Prob* > |*t*|" are now listed as "-" when they cannot be calculated. Previously, they were listed as "0.0000".

Changes were made in the calculation engine to the grid used when the minimum number of points to place between adjacent observed *x* values in the grid search is specified as greater than zero. The previous version of the program used a grid such that:

> if *X*(*j*), *j* = 1,..., *n* are the observed values of *X*,
>
> find X(k), the smallest observed value >= the current JoinPoint.
>
> allow the minimum next JoinPoint to be X(k+NBETW+1).

The current version allows the minimum next JoinPoint to be the first value in the grid after the specified number of observed *x* values. So, re-runs of previous JoinPoint models with the number of points between observations greater than zero could now give different results.

The options for "Graph Settings" have been enhanced to allow scaling of the X-axis.

The online help system has an updated format as well as updated information supporting version 2.6. The format has been changed from a one-pane help screen with pop-ups displaying information, to an HTML help format with a two-pane window. The left pane displays the contents or index while the right pane concurrently displays the selected help text.

You can now control the closing of the calculation engine's DOS window. Select **Preferences** from the **File** menu. A checkbox has been added that reads, "Automatically close DOS window when calculations complete". If you remove the check from this box, the DOS window will stay open after the calculation engine finishes processing. If you choose to have the window remain open, you will be able to view error messages provided by the calculation engine.

Version 2.5.2

Version 2.5.2 was released on November 16, 2000. No changes were made to the calculation engine. The program was modified to prevent a "domain error" when estimating the execution time. This error would not occur in a typical JoinPoint analysis, it would only happen during the analysis of extremely small data sets.

Version 2.5.1

Version 2.5.1 was released on November 10, 2000. No changes were made to the calculation engine. The default value for "The number of randomly permuted data sets for permutation tests" was changed to 4499. The default value is the value shown on the **Details** tab when you first execute JoinPoint or when you select **New** from the **File** menu.

This has no impact on analyses that you perform with saved parameter files. When you open a parameter file, the number of randomly permuted data sets will be set to the value stored in the file. If you want to use 4499 in previous analyses then you need to open the parameter file, go to the **Details** tab and change the value if necessary.

Version 2.5

Version 2.5 was released on July 18, 2000. The changes made to the program are described below.

In previous versions, the random number generator was seeded once at the start of the analysis. In this new version, the generator is re-seeded, using the seeds defined on the **Details** tab, for each by-group. This guarantees that the output from a single run of the program using by-variables will produce the same results for each level of the by-variable as the set of outputs produced by separately inputting each level of the by-variables into the program. (For a general discussion on fixing the seeds, see FAQ #5 on the JoinPoint web site). Because of this change, results from the new version that use by-variables may not match the results from the previous version.

> Recommendation:
>
> Although the default number of permutations is still 999, for greater consistency in the *p*-values obtained if one were to change the seed for each run, we now suggest running the program for least 4499 permutations. In most instances the JoinPoint Regression Program is run to select models with between 0 and 3 JoinPoints, which entails 3 statistical tests each conducted at the Bonferroni adjusted cutoff *p*-value of .05/3 = .0167. The value 4499 was chosen so that if you obtained a *p*-value of .0167 using one seed with 4499 permutations, then, assuming the number of possible permutations is large, the complete run using all possible permutations would have approximately a 99% chance of the *p*-value being between .0120 and .0220 (length of confidence interval = .0100), and approximately a 95% chance of the *p*-value being between .0129 and .0206 (length of confidence interval = .0077). Selection of the number of permutations selected by the user is a trade off between computer time and consistency of the *p*-values obtained. After an initial trial period we may change the default number of permutations to 4499.

A correction was made in calculating the weights when the Heteroscedastic Errors Option is set to "Poisson with Rates". This change has no impact on analyses performed using other option settings.

It was determined that the Fortran calculation engine can only handle a maximum of 200 formats per by-variable. Each format label must be 41 characters or less. The user interface was modified to ensure that these requirements are met. The format labels are truncated if necessary.

The program verifies that there is a large enough number of observations in the data to meet the specifications on the **JoinPoints** tab (number of JoinPoints, minimum number of observations from a JoinPoint to either end of the data, etc.). Previous versions of the program mistakenly require 1 more observation than necessary.

If a parameter file is saved or executed and there are existing report and output files, the user is told that those files will be deleted and asked if it is OK to overwrite the output files. The previous version of the program deletes the JPR and JPO files regardless of the user's response. This was corrected in version 2.5.

Previous versions of JoinPoint rounded input data to 6 decimal places. Version 2.5 maintains the precision of the input data.

Version 2.4

Version 2.4 was released on May 26, 2000. This version was developed to correct two problems in the user's interface. These changes have no impact on the calculations. The changes are described below.

The program verifies that there are enough observations in the input data to meet the specifications on the **JoinPoints** tab (number of JoinPoints, minimum number of observations from a JoinPoint to either end of the data, etc.). Previous versions of the program mistakenly required 1 more observation than necessary.

If a parameter file is saved or executed and there are existing report and output files, the program puts up the message: "The report and output files exist. Is it OK to overwrite these files?" The program mistakenly deletes the report and output files regardless of the response.

Version 2.3

Version 2.3 was released on March 1, 2000. This version was developed to correct a problem that occurred when the program eliminated sub-groups containing a zero response rate. In Version 2.2, an extra sub-group was removed from the analysis for every group that needed to be removed. This was corrected in Version 2.3. In all versions, a warning message was displayed if a zero response rate was encountered.

Note: the version information within the program (in the About dialog in the **Help** menu and in the JoinPoint report) was not updated on March 1, 2000. A new executable was made on March 10, 2000 with the updated version number.

Version 2.2

Version 2.2 was released in February, 2000. This version incorporated corrections and enhancements made in response to findings during the beta test period for Version 2.1 (beta). In addition, one change was made to the program as described below.

The input value can not have a value of zero for the response variable if the Heteroscedastic Errors Option is set to either "Input Variance of Response" or "Input Standard Dev of Response". In previous versions, the JoinPoint Regression Program would check the entire input data file to see if any response rate equaled zero. If so, the analysis was not performed at all. Version 2.2 was developed so that stratified groups that contain a record with a zero response rate are thrown out of the analysis, but all other groups are analyzed.

Version 2.1 (beta)

Version 2.1 was released in December, 1999. This version was developed to incorporate several revisions to the user interface and JoinPoint analysis report. The following changes were made:

The setting of the Heteroscedastic Errors option was changed in the JoinPoint user interface.

The time required to perform the analysis is estimated by the program. The program displays a message providing the estimate if the time is greater than one minute.

The JoinPoint report was revised to identify the stratified groups using variable labels instead of values.

The format of many statistics in the JoinPoint report was changed from scientific notation to decimal notation.

Confidence intervals that were in the report in previous versions were removed.

In previous versions, the user could set the names of the report and output files. In this version, the names of these files are determined by the program and are based on the name of the parameter file.

Version 2.0

Version 2.0 was released in June, 1999. This version was developed to incorporate several enhancements including the full implementation of by-variables and loading the description, variable positions, by-variable formats, and other file format information from SEER*Stat Export dictionaries.

In addition, the following changes were made to the JoinPoint Regression Program:

Autocorrelated errors option: in previous versions the program estimated the autocorrelation; now the user must specify an autocorrelation.

Minimum number of observations between two JoinPoints: in previous version the default was 1 and the minimum allowable value was 1; now the default is 2 and the minimum allowable value is 0.

The program now prints a 99% confidence interval for the permutation test *p*-value.

The output data file has been revised to make it easier to import the data into graphics software such as PowerPoint, Excel, or Harvard Graphics.

Column headings were added to the output file that contains the graph data. The variable names appear between each set of data. The JoinPoint models are identified as JP0, JP1, etc. The selected JoinPoint model is identified with an asterisk (JP0* or JP1*).

Version 1.1b

Version 1.1b was released in October, 1998. This version was developed to correct a problem in graphs created by the JoinPoint Regression Program. In previous versions, year labels shown on the X-axis were incorrect if there were gaps within the year range. For example, if the data contained the years 1973-1983, 1988-1995 then the X-Axis labels would be incorrect. This problem was corrected in Version 1.1b.

Version 1.1a

Please note that the changes implemented in this version affect the results for some jobs.

Version 1.1a was released on June 19, 1998. This version was developed to implement the following changes to the calculation engine:

A new algorithm for sequential testing of the number of JoinPoints was implemented. Under the old algorithm, a Type I error was defined to be the selection of more than zero JoinPoints when the true number of JoinPoints is zero. Under the new algorithm, a Type I error is defined to be the selection of too many JoinPoints (regardless of the true number of JoinPoints). The new algorithm is defined as follows:

set Null = minimum number of JoinPoints, ALT = maximum number of JoinPoints

test NULL JoinPoints vs. ALT JoinPoints.

If NULL is rejected, then set NULL = NULL + 1; otherwise set ALT =ALT + 1.

Iterate steps b) and c) until NULL = ALT.

Estimates of the intercept and slope in each segment, and their standard errors, are included in the report file.

The minimum number of observations between JoinPoints can equal zero. If the chosen model has 0 (or 1) observations between two JoinPoints, the program prints the warning, "Information matrix is singular, generalized inverse used", sets the standard errors of the slope and intercept (or slope only) in that segment equal to zero and sets the corresponding *p*-values for the Wald tests equal to 1. The permutation test of the number of JoinPoints is still performed.

Version 1.0b

Version 1.0b was released in March, 1998. This version was developed to correct a problem in the JoinPoint Regression Program interface. In Version 1.0a, the Input Data File was being deleted when a JoinPoint session was executed. This problem was corrected in Version 1.0b.

Version 1.0a

Version 1.0a was released in February, 1998. This was the first version of the JoinPoint Regression Program made available to users.

JoinPoint: Frequently Asked Questions

Program Changes

What are the most recent changes to the JoinPoint program?

Answer: Program changes for each version of JoinPoint are described in the Program Changes chapter of the help system.

How can I replicate previous results with a newer version of JoinPoint?

Answer: Because of changes in statistical methodology, computational issues, and the random number generator from one version to the next, the same version of the program should be used again to exactly replicate results. If you would like prior versions of JoinPoint to replicate previous results, please request them by sending an email to JoinPoint Technical Support.

Model Estimation and Selection Methods

Why doesn't the JoinPoint program give me the best possible fit? I can see other models with more JoinPoints that would fit better. Exactly how does the program decide which tests to perform and which JoinPoint model is the final model?

Answer: As with many statistical models, if you add more parameters you get a better fit. The same is true with JoinPoint models. What the program does is to try to choose the smallest number of JoinPoints such that if one more JoinPoint is added, the improvement is not statistically significant. Thus, in the final model you may interpret each of the JoinPoints and its corresponding changes in trend as significant.

JoinPoint selects the final model using two different methods: Permutation Test and Bayesian Information Criterion (BIC). First, the user specifies MIN as the minimum number of JoinPoints and MAX as the maximum number of JoinPoints on the **JoinPoint** tab.

Then the program uses a sequence of permutation tests to select the final model. Each one of the permutation tests performs a test of the null hypothesis H0: number of JoinPoints = *ka* against the alternative Ha: number of JoinPoints=kb. The procedure begins with *ka* = MIN and *kb* = MAX. If the null is rejected, then increase *ka* by 1; otherwise, decrease *kb* by 1. The procedure continues until *ka* = *kb* and the final value of $\hat{k} = k_a = k_b$ is the selected number of JoinPoints.

The second method is based on the Bayesian Information Criterion (BIC). The value of BIC is the loglikelihood value penalized by the cost of extra parameters. The model with the minimum value of BIC is selected as the optimal model.

Describe the permutation test used here.

Answer: In this program, the permutation test is used repeatedly for testing between two different JoinPoint models, a simpler model with fewer JoinPoints called the null model, and a more complicated model called the alternative model. The alternative model fits better because it is more complicated.

The question for the test is: does it fit much better than would be expected by chance. To test this statistically, we calculate a ratio, SSEN/SSEA, where SSEN is the sum of squared errors (SSE) from the null model and SSEA is the SSE from the alternative model. Values of the ratio close to 1 mean that the alternative is not much better than the null model, while larger values mean that the alternative is much better.

In order to decide how much larger a ratio needs to be to be statistically significant, we use the permutation method. In this method we randomly permute (that is, shuffle) the errors (also called the residuals) from the null model and add them back onto the modeled values from the null model to create a permutation data set. Then we calculate the ratio for the permutation data set.

If the true model was the null model we would expect that about half of the ratios calculated from the permutation data set would be less than the original one.

If the true model was the alternative model, we would expect that after permuting the errors most of the new ratios would be less than the original ratio.

In other words the permuted data set would look less like the alternative model than the original data. So we reject the null model (or null hypothesis) if less than a certain proportion of the ratios are greater than or equal to the original ratio.

For more specific details, see Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation Tests for JoinPoint Regression with Applications to Cancer Rates. Stat Med 2000;19:335-351. To request a reprint, email Mr. Reggie Taborn for a copy at: mailto:tabornr@mail,nih.gov

How many permuted data sets should I use?

Answer: For greater consistency in the p-values obtained if one were to change the seed for each run, we suggest running the program for at least 4499 permutations. For this reason, the default number of permutations is now 4499 in the current version of JoinPoint (it was 999 in previous versions). In most instances the JoinPoint program is run to select models with between 0 and 3 JoinPoints, which entails 3 statistical tests each conducted at the Bonferonni adjusted cutoff p-value of .05/3 = .0167. The value 4499 was chosen so that if you obtained a p-value of .0167 using one seed with 4499 permutations, then, assuming the number of possible permutations is large, the complete run using all possible permutations would have approximately a 99% chance of the p-value being between .0120 and .0220 (length of confidence interval = .0100), and approximately a 95% chance of the p-value being between .0129 and .0206 (length of confidence interval = .0077). Choice of the number of permutations selected by the user is a tradeoff between computer time and consistency of the p-values obtained.

The JoinPoint program uses Monte Carlo simulation to calculate p-values for a series of permutation tests. See Permutation Test Details for details on the permutation tests performed and the data sets used to perform one permutation test in the series.

Here we discuss the implications of the choice of the number of permutation data sets, say N. The program runs faster with smaller values of N, but it gives better power with larger values of N. In addition, a larger N reduces the probability that another analysis of the same data might get a different answer when run with different random number generator seeds. (Computer programs produce pseudo-random numbers through algorithms that mimic randomness, which we use to shuffle or permute the errors. The algorithms use a seed or seeds to start the algorithm. These seeds can be used to produce repeatable pseudo-random numbers.)

The problem of two analyses obtaining different answers from the same data is addressed by this program by specifying default random number generator seeds. Thus, as long as no parameters are changed (including the random number generator seed and N), repeats of the analyses will produce the same results. Otherwise, two runs of the same analysis except with different seeds could get different answers.

To get an idea how results would change for someone using different random number generator seeds, we list some confidence intervals for p-values below. For example, with N=999 Monte Carlo samples if you obtained a p-value of .04 from the program there is an approximately 99% chance that another researcher repeating the analysis with N very large (i.e., an ideal situation with N -> infinity) would obtain a p-value between .025 and .0577.

N=99

lower 99% ci	estimate	upper 99% ci
0.0000	0.01	0.0521
0.0034	0.04	0.1065
0.0069	0.05	0.1218
0.0111	0.06	0.1364
0.0325	0.10	0.1910
0.02702	0.40	0.5281

N=999

lower 99% ci	estimate	upper 99% ci
0.0031	0.01	0.0199
0.0250	0.04	0.0577
0.0331	0.05	0.0694
0.0415	0.06	0.0810
0.0762	0.10	0.1259
0.3595	0.40	0.4402

N=9999

lower 99% ci	estimate	upper 99% ci
0.0075	0.01	0.0127
0.0350	0.04	0.0452
0.0445	0.05	0.0558
0.0540	0.06	0.0663
0.0923	0.10	0.1079
0.3873	0.40	0.4127

N=99999

lower 99% ci	estimate	upper 99% ci
0.0092	0.01	0.0108
0.0384	0.04	0.0416
0.0482	0.05	0.0518
0.0581	0.06	0.0620
0.0976	0.10	0.1025
0.3960	0.40	0.4040

What does the *p*-value mean for JoinPoint?

Answer: The JoinPoint Regression Program performs a series of hypothesis tests that test the null hypothesis of *ka* JoinPoints against the alternative hypothesis of *kb* JoinPoints, where *ka* and *kb* change for each hypothesis test. Each *p*-value corresponds to this type of test. The *p*-value is an estimate of the probability, under the assumption that there are only *ka* JoinPoints, of observing data that look more like a JoinPoint model with greater than *kb* JoinPoints than the data that we have in fact observed.

As the permutation test is a randomization test, which depends on the random number generator. For greater consistency in the permutation test *p*-values obtained if one were to change the seed for each run, we suggest running the program for at least 4499 permutations. For this reason, the default number of permutations is now 4499 in the current version of the JoinPoint Regression Program. Choice of the number of permutations selected by the user is a tradeoff between computer time and consistency of the values obtained.

Why are new significance levels being used in Version 3.0 and succeeding releases?

Answer: The JoinPoint software uses a series of permutation tests to determine the number of JoinPoints. Prior to Version 3.0, the software used the Bonferroni adjustment to control the error probability of each of the multiple tests. The Bonferroni adjustment has been shown to be conservative and the procedure tends to select fewer JoinPoints than it should. The procedure with new significance levels controls the over-fitting probability and still maintains the correct alpha level. It is superior to the traditional Bonferroni adjustment.

Details: The Bonferroni adjustment is conservative because the actual overall significance level is usually less than the nominal level α. Starting with Version 3.0, the new adjustment procedure controls the overall over-fitting error probabilities,

$P(k > K_a \mid k = K_a)$, $K_a = \text{MIN...MAX}$,

under α. Let k denote the number of JoinPoints and $\alpha(k_a; k_b)$ be the significance level of each individual test H0 : $k = K_a$ vs. H1 : $k = K_b$. The new procedure set $\alpha(K_a; K_b) = \alpha/(\text{MAX} - K_a)$. Notice that the individual significance level depends on the number of JoinPoints K_a under the null. Consider an example where MIN = 0 and MAX= 4. The new procedure has the following properties:

$P(k > 0 \mid k = 0) = \_(0,4) + \_(0,3) + \_(0,2) + \_(0,1);$

$P(k > 1 \mid k = 1) = \_(1,4) + \_(1,3) + \_(1,2);$

$P(k > 2 \mid k = 2) = \_(2,4) + \_(2,3);$

$P(k > 3 \mid k = 3) = \_(3,4).$

If we like to bound these over-fitting probabilities by α, then we can assign different values for each (k_a; k_b). That means, we can achieve a better power by setting

$\alpha(0,4) = \alpha(0,3) = \alpha(0,2) = \alpha(0,1) = \alpha/ 4;$

$\alpha (1,4) = \alpha (1,3) = \alpha (1,2) = \alpha / 3;$

$\alpha (2,4) = \alpha (2,3) = \alpha / 2;$

$\alpha (3,4) = \alpha$

What is the Bayesian Information Criterion method for selecting the best model?

Answer: Permutation test and BIC are two different approaches to selecting the optimal number of JoinPoints. The Permutation test approach uses a sequence of permutation tests to determine the true number of JoinPoint. The PT approach controls the error probability of selecting the wrong model at a certain level (i.e. 0.05), whereas the BIC approach finds the model with the best fit by penalizing the cost of extra parameters. The models picked by BIC tend to fit the data well but are less parsimonious. The applications have shown that the PT approach worked well for cancer incidence and mortality data.

The equation for computing the BIC for a k-JoinPoint model is:

BIC(k) = ln{SSE(k)/#Obs(k)} + {#Parm(k) /#Obs} * ln(#Obs),

where SSE is the sum of squared errors of the k-JoinPoint regression model, #Parm(k)=2*(k+1) is the

number of parameters of the k-JoinPoint model and #Obs is the number of observations.

The k-JoinPoint model with the minimum value of BIC(k) is selected as the final model.

Why should I use the Permutation Test (PT) when Bayesian Information Criterion (BIC) is so much faster?

Answer: Permutation test and BIC are two different approaches to selecting the optimal number of JoinPoints. The Permutation test approach uses a sequence of permutation tests to determine the true number of JoinPoint. The PT approach controls the error probability of selecting the wrong model at a certain level (i.e. 0.05), whereas the BIC approach finds the model with the best fit by penalizing the cost of extra parameters. The models picked by BIC tend to fit the data well but are less parsimonious. The applications have shown that the PT approach worked well for cancer incidence and mortality data.

Should I use the linear or log-linear model?

Answer: The linear or log-linear model can be chosen depending on how linear the observed rates or the logarithm of the observed rates are over time. In order to check the goodness of the chosen model, a user can test for normality of the residuals obtained under the linear or the log-linear fit. Select a model whose residual analysis indicates a better fit, regarding the model assumptions of normality, linearity, equal variance, and independence. One reason for using a log transformation for cancer rates is that they arise from a Poisson distribution which is skewed especially when the cancer is rare or the rates come from a small population. The log transformation is a standard way to make this skewed distribution approximate a normal distribution. Rates for common cancers or which come from a large population can be approximated as arising from a normal distribution without a transformation.

One motivation for using the log-linear model for cancer rates regardless if they are rare or not is the ease of interpretation. Under a log-linear model the rates change at a constant percent per year (i.e. a fixed annual percent change - APC), while for a linear model the rates change at a constant fixed amount per year. When comparing trends across age group or across cancer sites where the rates are very different, the advantage of a log-linear model is that the APC is a metric which makes sense to compare across widely different scales. For example, a rare cancer and a common cancer may change at the same annual percent per year, but it is highly unlikely that they would change at the same fixed amount per year (e.g. if the rates were declining the rare cancer rate would quickly become negative!).

When running the test for pairwise differences with the Coincident option, should I use a weighted or unweighted model?

Answer: If the weights for the two cohorts are fairly different, the combined model will be heavily influenced by the cohort with the larger weight, and the results may not be as expected since the combined fit will closely mimic the level and number and location of JoinPoints for that individual cohort. In this case the statistical algorithm is appropriately weighting the series that is more reliable. In some cases, however, it may be more appropriate to weight each series equally, even though one series has a larger variance.

For example, it may be appropriate to weight two racial/ethnic groups equally if the goal is to find the best fit for two groups ignoring the fact that one series is more reliable than the other. In such a case, running the unweighted model using the Heteroscedastic Errors Option of "Constant Variance" may be more appropriate. These same considerations are less obvious but can be relevant when fitting a parallel model. In this case, level is not an issue (since each cohort has its own level), but the fit of the number and location of JoinPoints will be heavily influenced by the larger cohort, unless an unweighted analysis is used. There is no "correct" answer, but careful consideration should be given to the overall purpose of the analysis in these situations.

For more details about the comparability or pairwise differences test, see:

http://surveillance.cancer.gov/JoinPoint/comparabilitytest.html.

Does the test of parallelism of two series require the series to be uncorrelated?

Answer: The permutation test for parallelism assumes exchangeability in the pair of residuals at each time period. When errors of the two series are correlated, this assumption is violated, so the permutation pvalue will not be accurate. Errors in series are correlated, for example, when the two series come from the same population or sub-population measured from the same survey (e.g. obesity and diabetes for white males from the same survey) or the same cancer registry (e.g. white male colorectal and prostate cancer rates from the same registry). Errors in series may be considered uncorrelated when the observations in the two series come from different samples of the same population (e.g. obesity for white males measured from one national survey and diabetes for white males measured from a different national survey).

One way to test for parallelism in two series with correlated errors is to compute a new time series, the difference between the original two series. The series are parallel if the difference time series has a constant mean. One way this can be checked is by fitting a simple linear regression and testing whether the slope coefficient is zero.

Can the JoinPoint software estimate the autocorrelation parameter?

Answer: No, the software cannot estimate the autocorrelation parameter. In our paper in Statistics in Medicine (Kim, et al, 2000, 335-351) we outlined a method for estimating the autocorrelation parameter (see section 2.3). We found in our simulations on table IV of that paper that adjusting for autocorrelation was helpful in maintaining proper size of the tests of JoinPoints when there was large autocorrelation. We also found that if there was no autocorrelation that the adjustment seriously affected the power of the test to detect JoinPoints. For example we see in Table IV (b) with phi=0, the power goes from 90% to 68%. Thus we chose not to include an option in the JoinPoint software for estimating autocorrelation as done in the paper. This is to avoid this serious loss of power when there is no autocorrelation present. Instead we allow the user to input values for autocorrelation as a type of sensitivity analysis. The way this option would be used is as follows:

> Fit the model with the uncorrelated errors option.
>
> If the user suspects that there is positive autocorrelation in the data, then repeat the analysis trying several values of the autocorrelation parameter, say for example .1, .2, and .3. If the results are very similar with different values of the autocorrelation parameter, then the user knows their results will still hold if there is autocorrelation present. If the results change as the autocorrelation parameter changes then the user may end up presenting the series of results, to show how the results depend on different assumptions about the autocorrelation.
>
> If the user suspects negative autocorrelation, there is need to do any further analysis (see Kim, et al., 2000).

Applications of JoinPoint

I used the program on SEER data and do not get the same answers as in the SEER Cancer Statistics Review. Why?

Answer: The SEER Cancer Statistics Review (CSR) uses the same JoinPoint Regression Program. Their analysis uses the "Input Standard Error of Response" setting for the Heteroscedastic Errors Option using the standard error of the rate as calculated by SEER*Stat. Using these options, and the current default number of permutations (4499), if the JoinPoint Regression Program chooses the model with 0 JoinPoints, the annual percentage rate change will agree with the calculation of this value given in the CSR.

How do I import JoinPoint's graphs into a graphics software package such as Excel, Harvard Graphics, or PowerPoint?

Answer: The graphs themselves, or the data to produce the graphs, can be exported from JoinPoint into a format which can be used by other software packages.

> Use **Output > Export Graph...** to save a bitmap (.bmp) graphics file which contains an image of the currently displayed graph.

> Use **Output > Export as Text > Data...** to save a text file which contains the coordinates (X, Yobserved, Y-fitted) used for creating the graphs. This file can be imported into other graphics software packages.

When I fit a JoinPoint model to find the rate trend, why doesn't the APC for one of the segments match the APC found by SEER*Stat for the corresponding segment?

Answer: The JoinPoint model assumes that the trend is continuous at the JoinPoint, whereas the APC for the corresponding segment calculated by SEER*Stat does not impose the continuity constraint with the consecutive segment. Hence, when there are one or more JoinPoints, the APCs from JoinPoint and from SEER*Stat are different, but they should be close to each other. When there are zero JoinPoints, the APCs from JoinPoint and SEER*Stat will be the same, except for differences due to the precision level of the data.

Sometimes, the change in the slopes between two segments is not statistically significant (*p*value>0.05) from the *t*-test, but there is a JoinPoint between the two segments or vice versa. Why?

Answer: The JoinPoint program shows the estimates of the regression coefficients, i.e., intercepts and slopes, and the changes of the slopes. The *p*-values of the slope changes are calculated from the *t*-test based on asymptotic normality. Based on a *p*-value greater than 0.05, one might say the two slopes are not statistically different and hence conclude that the two segments are the same. But using the *p*-value from a *t*-test is not as accurate as that from the permutation test, since the *t*-test is an asymptotic test and the variances are calculated using the information matrix conditional on the estimated JoinPoints without imposing the continuity constraint and omitting the offending observations. The software (with the permutation test procedure) does not require the asymptotic normality and maintains the correct Type I error probability level and hence the number of JoinPoints determined by the software is more reliable.

Sometimes, the APC for one segment is significantly different from zero, when an extra JoinPoint in the segment is determined by the JoinPoint software, neither APCs for the two consecutive segments are significant. Why?

Answer: First, the test of APC is based on asymptotic *t*-test and the number of JoinPoints is determined by permutation test, which is more reliable. When a segment is divided into two segments by a JoinPoint, there are less data points in each segment, which usually decreases the power of each test. An omnibus test, which combines both segments, is a more powerful test to detect the overall trend for both segments.

Can I do JoinPoint regressions where the time intervals are not equally spaced or the time points are not integer?

Answer: Yes. The JoinPoint software since Version 3.0 allows unequally spaced time intervals and non-integer time points.

How can JoinPoint be used to model proportions or percents?

Answer: JoinPoint can be used to modeling proportions or percents, but usually one would want to incorporate the standard errors of the proportions into the analysis. For consistency, assume that percents are transformed to proportions before analysis. For confidence intervals and p-values, JoinPoint assumes that the data being analyzed arises from a normal distribution. If n_i is the sample size for the *ith* observation and is large enough that both n_i p_i and n_i(1-p_i) are larger than or equal to 10, then the proportions can be considered as coming from an approximately normal distribution with standard error equal to $\sqrt{p_i\,(1-\,p_i)\,l\,n_i}$. If the proportions arise from a complex survey, then the standard error from a complex survey analysis statistical package (e.g. SUDAAN) can be used. If is not large enough for a normal approximation (that is, either n_i p_i or n_i(1-p_i) or both are smaller than 10), then the distribution may be skewed, and JoinPoint results based on asymptotic normality may not be accurate.

If n_i is not available, but all the proportions are in approximately the same range, and the sample sizes are known to be approximately the same, then it may not be necessary to enter the standard errors (i.e. in this case all the standard errors are approximately the same, which is what is the implicit assumption in JoinPoint if no standard errors are entered).

How does the JoinPoint software deal with zeros in the dependent variable?

Answer: It depends on which Model Type and Heteroscedastic Errors Option are selected:

> If one of the Poisson models are selected, 0.5 is added to • each of the counts.
> If the linear model option is selected, then the data is analyzed normally.
> If the log-linear model option is selected, then the by-group with a zero is dropped from the analysis.

Can use JoinPoint to project or extrapolate into the future?

Answer: While the JoinPoint program itself does not do projections/extrapolations, it is not unreasonable to use JoinPoint to project a few years into the future. To do so, one would need to extrapolate the final segment, and compute the standard errors of the projection, using the betas displayed in the output. The JoinPoint Team has had some success using JoinPoint for projections (validated against actual values when the data is rolled back several years so the projected values can be validated against actual values). However, a projection using JoinPoint may or may not be more accurate than other methods (e.g. time series methods), depending on the situation. Long term projections using JoinPoint are not advised. Also, AAPCs, since they are summary measures over an interval, are not as appropriate to use for projections as is the APC of the final segment.

A reference to an article which used JoinPoint for projections is:

Pickle, L., Hao, Y., Jemal, A., et al. (2007) "A new method of estimating United States and state-level cancer incidence counts for the current calendar year" CA: A cancer journal of Clinicians 57:30-42.