Invited article

Outbreak, Surveillance, Investigation & Response (OSIR) Journal
ISSN 2651-1061 (Online)
https://he02.tci-thaijo.org/index.php/OSIR
Email: osireditor@osirjournal.net

March 2026
Volume 19, Issue 1
Article no. 281011

The Grammar of Science: "Knots" and "Pieces"

Jaranit Kaewkungwal*

Faculty of Tropical Medicine, Mahidol University, Thailand

*Corresponding author, email address: jaranitk@biophics.org

Received: 9 Apr 2025; Revised: 11 Dec 2025; Accepted: 19 Dec 2025
https://doi.org/10.59096/osir.v19i1.281011

When I started planning to write about piecewise regression, two words immediately came to mind: "knots" and "pieces." In the Oxford Learner's Dictionary, a knot is a join made by tying two pieces together, and a piece is simply part of something larger. That idea fits well with many problems in health research, where the relationship between an exposure and an outcome is rarely a single, simple pattern. Drug responses can change across dose levels, cognitive aging may shift after retirement, biological thresholds can affect disease detection, and policies or interventions can alter behavior over time. In situations like these, the key question is not whether a relationship exists, but when it changes. This paper introduces how piecewise linear regression works and discusses when it is most appropriate to use.

Core Concept of Piecewise Regression

Piecewise linear regression, also known as segmented or spline regression, models relationships that change at specific points by fitting connected linear segments to the data. It is useful when a single straight line cannot capture clear shifts in trend and is commonly applied around cutoff values or key events, such as interventions, laboratory thresholds, or disease transitions.[1–3] These points, called breakpoints or "knots," mark where the slope changes.[1] Knots may be identified visually or estimated from the data and should reflect meaningful real-world thresholds.[3,4] Once included in the model, they divide the relationship into linear segments or "pieces", allowing the trend to change while remaining smoothly connected.

Although piecewise and segmented regression are essentially the same modeling approach, spline regression shares a similar foundation but differs conceptually.[2,5–7] The primary aim of a piecewise model is to detect or test a changepoint. It fits separate linear segments with slope changes at predefined breakpoints, which are often substantively meaningful. This makes interpretation straightforward and emphasizes explicit threshold effects.[5–7] In contrast, spline regression is designed to flexibly model nonlinearity. It enforces smoothness at knots, which are typically chosen based on the data, to capture gradual changes in the relationship. Depending on the type of spline, the model may require explicitly specified knots, such as in regression splines, or may handle smoothness automatically without user-defined knots, as in some smoothing spline approaches. While splines are well-suited for smooth continuous patterns, their interpretation can be more complex.[8–10] As practical examples, piecewise regression models can be used to evaluate an intervention introduced at Month 6, to test whether the slope changes before versus after a policy implementation, or to assess outcome changes at a body mass index (BMI) threshold of 25 kg/m^2. In contrast, spline regression is more suitable for modeling age versus cognitive score across the lifespan, examining dose–response relationships, or capturing outcome changes driven by continuous biological processes.

In longitudinal studies such as randomized clinical trials or observational follow-up research, changes in the mean response over time are often not strictly linear; the data may appear linear over short intervals before shifting direction. Although higher-order polynomial models can approximate these patterns, they risk overfitting and may behave unstably at the boundaries, where small data changes can cause large swings at the extremes.[8,9,11,12] Piecewise linear regression may offer a more flexible and interpretable alternative for modeling such nonlinear trends. By fitting segmented linear models, common nonlinear patterns can be captured clearly and effectively.[2,10,13]

As shown in Figure 1, the coronavirus disease 2019 (COVID-19) data do not follow a single linear trend. A standard linear regression, therefore, provides poor predictions, as it fails to capture the rise–fall pattern. Although polynomial models may improve overall fit, their parameters are often difficult to interpret meaningfully. In contrast, piecewise regression offers a practical alternative by introducing a meaningful breakpoint, such as the implementation of preventive measures during the COVID-19 epidemic. This approach captures differences in case rates before and after the intervention, improving both model fit and interpretability.

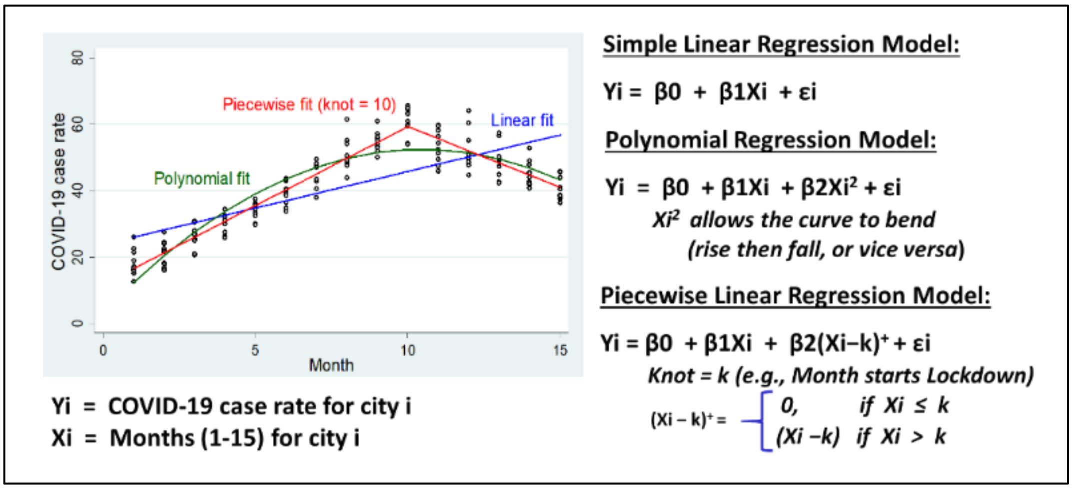

Figure 1. Model fit for COVID-19 case rate over time.

Figure 2 demonstrates that a single linear trend does not adequately describe the relationship between lisinopril dose and change in systolic blood pressure. A piecewise regression model with knots at 2.5, 10, and 20 mg/day substantially improves model fit. A restricted cubic spline model produces a smoother functional form, yet in this case yields very similar results. This similarity is likely because the knots were specified at the same dose cutoffs and the administered doses are discrete rather than continuously distributed. As a result, both models capture the underlying pattern in a comparable manner. If dose levels varied more continuously across a broader range, the spline approach would be expected to display a smoother curve and potentially highlight more gradual nonlinear trends. Overall, piecewise regression is preferable when the focus is on clear cutoffs and straightforward interpretation, whereas restricted cubic splines are better suited for modeling continuous, smoothly changing dose–response relationships.

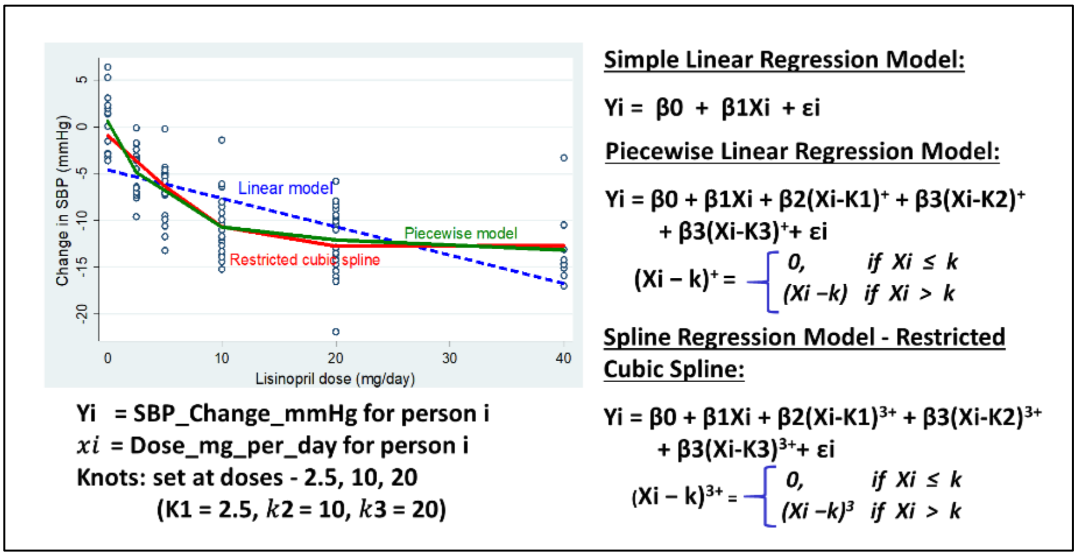

Figure 2. Model fit between lisinopril dose and systolic blood pressure change.

Statistical Foundations of Piecewise Models

Regression models are used to predict outcomes based on independent variables and to examine relationships between variables for forecasting purposes. Regression is a predictive analysis technique that explores the relationship between one or more independent variables (X) and a dependent variable (Y). The fundamental objective is to fit a mathematical function that describes how the outcome changes as the predictors vary.[11,13,14] In simple linear regression, the dependent variable Yi for observation i is predicted from an independent variable Xi and is expressed as:

$$Yi = \beta 0 + \beta 1Xi + \varepsilon i$$

where β0 is the intercept, β1 is the slope, and εi is the error term.

The coefficient β1, also referred to as the slope or dY/dX, represents the rate of change in the outcome Y with respect to the predictor X.[11,15] Thus, β1 indicates that for every one-unit increase in X, the expected value of Y changes by β1 units.

Piecewise linear regression builds on the idea that the relationship between variables may change across different ranges of the predictor. Instead of fitting a single straight line to all observations, piecewise linear regression uses multiple linear segments joined at specific breakpoints (knots).[3,15] These changes are modeled by fitting separate linear functions across different ranges of the predictor. Similar to simple linear regression, the model is typically estimated by minimizing the average difference between the predicted and observed values, usually through the least squares method.[5,7,10] The piecewise model is commonly formulated using interaction terms or dummy variables to ensure continuity at the connecting points between segments.[16] Once a knot is defined at breakpoint B, the model can be written as:

$$Yi = \beta 0 + \beta 1Xi + \beta 2(Xi - B)^{+} + \varepsilon i$$

where $(Xi - B)^{+} = \begin{cases} 0, & Xi > B \\ Xi - B, & Xi \leq B \end{cases}$

In this formulation, β0 = Intercept ; β1 = slope before the breakpoint; β1 + β2 = slope after the breakpoint. Thus, the slope changes once X exceeds the breakpoint B, while the function remains continuous.

One Breakpoint Piecewise Regression

To illustrate a one-knot model, the study examined the relationship between fasting glucose level (mg/dL) and BMI (kg/m²). As shown in Figure 3, a single simple linear regression line does not adequately capture the relationship across both non-obese and obese individuals. The association appears to shift around the conventional BMI cutoff of 25. To position the knot at BMI = 25, define: Xi = BMIi−25. This centers BMI at 25, such that: if BMI = 25 → Xi=0; if BMI <25 → Xi <0; if BMI >25 → Xi >0. Next, define a truncated (hinge) term: $(Xi)^{+}$ = max (0, BMIi−25). Thus: if BMI ≤25 → $(Xi)^{+}$ = 0; if BMI >25 → $(Xi)^{+}$ = BMIi−25. The corresponding piecewise regression model can be written as: Glucosei = β0 + β1Xi + β2$(Xi - 25)^{+}$ + εi. This specification allows the model to estimate different slopes before and after the cutoff at BMI =25, while maintaining continuity at the knot with one slope for individuals with BMI ≤25 and a different slope for individuals with BMI >25. As a result, the piecewise regression model provides a more flexible and better-fitting representation of the relationship between BMI and fasting glucose compared to a single straight line.

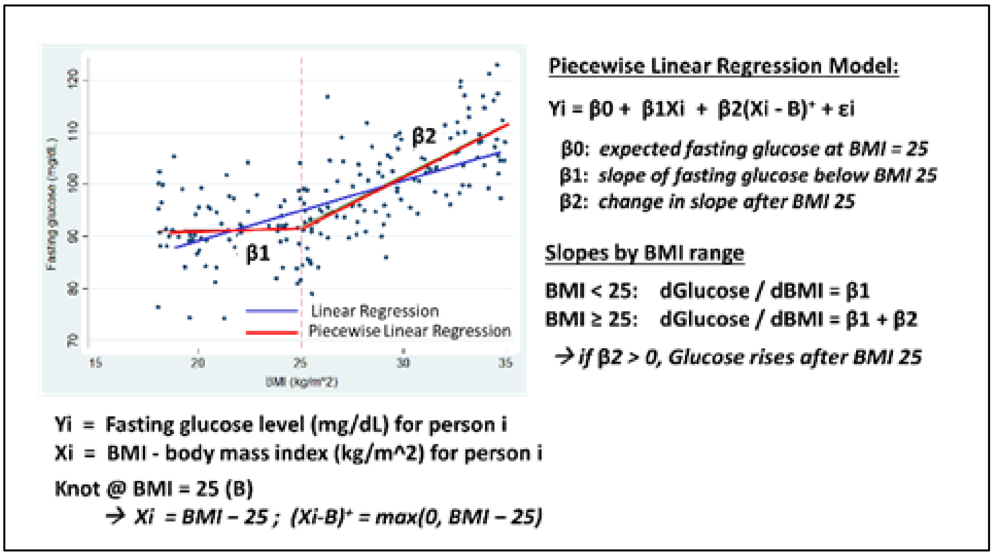

Figure 3. One-knot model for relationship between fasting glucose and BMI.

Suppose the result is:

$$\text{Glucose}_i = 70 + 1.2\text{BMI}_i + 2.5(\text{BMI}_i - 25)^+ + \varepsilon_i$$

where $(\text{BMIi} - 25)^+ = \begin{cases} 0, & \text{BMIi} \leq 25 \\ \text{BMIi} - 25, & \text{BMIi} > 25 \end{cases}$

Interpretation:

- Slope before BMI =25 (1.2) → For individuals with BMI ≤25, each 1-unit increase in BMI is associated with an average 1.2 mg/dL increase in glucose.
- Slope after BMI =25 → For BMI >25, each 1-unit increase in BMI is associated with an average 1.2 + 2.5 = 3.7 mg/dL increase in glucose.

Two-Breakpoints Piecewise Regression

To illustrate a two-knot model, the study examined the relationship between cognitive score and age. The knots were predefined at ages 60 and 75, corresponding to three age groups: 40–60, 61–75, and 76–90. Age was parameterized as Xi = Age−60, with truncated terms $(Xi−B1)^+$ = max (0, Age−60) and $(Xi−B2)^+$ = max (0, Age−75). Figure 4 illustrates the structure of the piecewise regression model for this scenario.

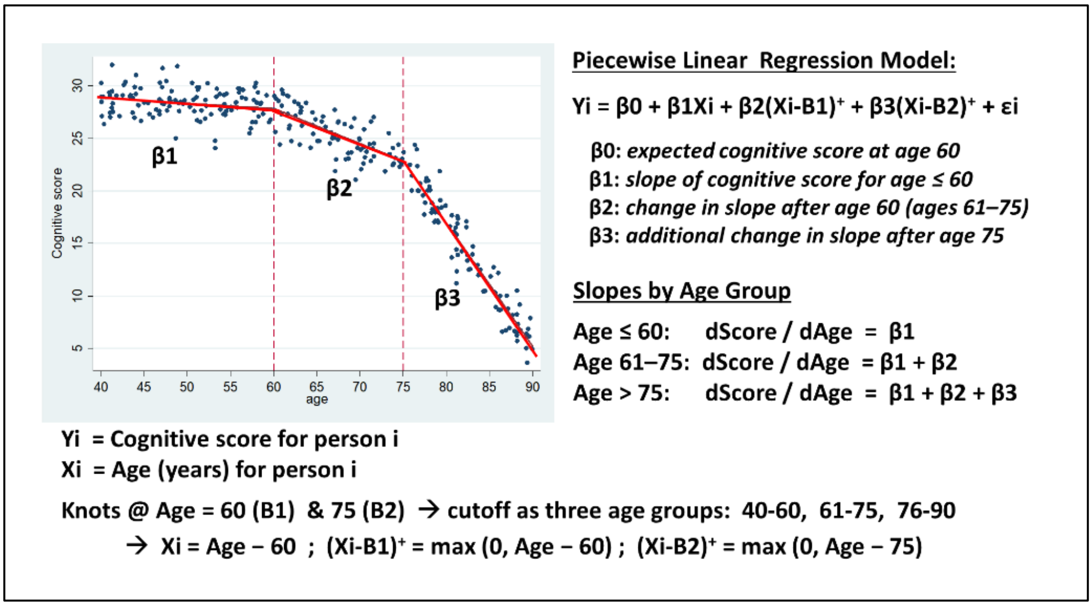

Figure 4. Two-knots model for relationship between cognitive score and age.

Suppose the result is: $\text{CognitiveScore}_i = 125 - 0.3\,\text{Age}_i - 0.8\,(\text{Age}_i - 60)^+ - 0.5\,(\text{Age}_i - 75)^+ + \varepsilon_i$

Interpretation:

- Before age 60: Slope = −0.3 → Cognitive score decreases by 0.3 points per year.
- Between ages 61–75: Slope = −0.3−0.8 = −1.1 → Cognitive score decreases by 1.1 points per year. (Age 60 represents a change point where decline accelerates.)
- After age 75: Slope = −0.3−0.8−0.5 = −1.6 → Cognitive score decreases by 1.6 points per year. (Age 75 represents a second increase in the rate of decline.)

Two-Breakpoints Piecewise Regression with Interaction between Groups

Comparing means before and after an intervention is a common application of piecewise analysis. This basic shift model can be extended by incorporating a time trend to capture changes over time in each group. In a pre–post design, separate linear trends are estimated for the period before the intervention and for the period during or after it. The model can also compare trends between intervention and non-intervention groups.[6,17] In the example shown in Figure 5, the outcome is the number of vaccinated children in each city. As this is a count variable, it may not satisfy the normality assumption required for linear regression. In such cases, Poisson or binomial regression models are more appropriate, as they are specifically designed for modeling counts and proportions. It should be noted that piecewise regression is not limited to linear models. The same segmented structure can be applied to other types of regression models, including generalized linear models, multilevel models, and mixed-effects models.[10,18]

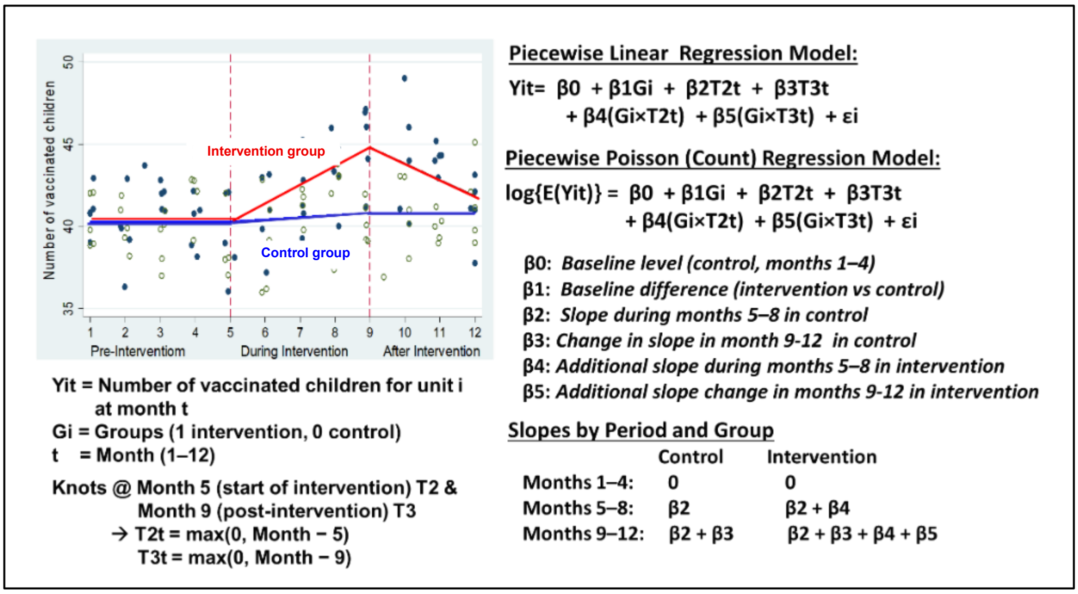

Figure 5. Two-knots model for comparing between groups over three periods.

An example is a study comparing two groups (intervention vs. control), where the outcome was the number of vaccinated children measured across three periods: the pre-intervention period (months 1–5), the implementation period (months 6–9), and the post-intervention period (months 10–12). To model changes across these phases, two knots were specified: Month 5 (denoted as T2) and Month 9 (denoted as T3). These knots represent the transition points between study phases. The spline terms were defined as: T2 = max (0, Month−5) and T3 = max (0, Month−9). This specification allows the model to estimate: one time trend during months 1–5, a different trend during months 6–9, and another trend during months 10–12.

Suppose the result is: Vaccination ratei = 50 + 2.0 Montht + 8 Groupi + 3.5 T2t + 4.0 T3t + 2.5 (Groupi×T2t) + 3.0 (Groupi×T3t) + εi

Interpretation:

- Intercept (50) → At Month 0 (baseline) in the control group, the expected number of vaccinated children is 50 children
- Time trend before intervention (Months 1–5) Coefficient of Month = 2.0 → In the control group, vaccination increases by 2 children per month during the pre-intervention period.
- Group effect (8) → At baseline, intervention cities have 8 more vaccinated children than control cities.
- Change during implementation (Months 6–9) T2t = max (0, Month−5) → Control group slope increases by 3.5 additional children per month after Month 5. Intervention group gains an extra 2.5 children per month beyond the control group during this period.
- Change after intervention (Months 10–12) T3t = max (0, Month−9) → Control group slope increases by 4 additional children per month after Month 9. Intervention group gains an extra 3 children per month beyond the control group in this final period.
- Overall Meaning → Before Month 5: gradual increase in vaccination, Months 6–9: stronger upward trend; Months 10–12: further acceleration; Intervention cities show larger increases compared to control cities during and after implementation.

Selecting and Confirming Knots

Although piecewise regression is conceptually simple, identifying the optimal breakpoints remains challenging.[5,15,19] The method balances two competing goals: minimizing error so that each segment fits the data closely, and using as few segments as possible to avoid overfitting.[2,5,15] In practice, breakpoint locations are selected to minimize the overall squared error, providing good fit within segments without unnecessary complexity.[2,3] Several recommendations exist for choosing knots in piecewise regression.[20–22] In health research, knots may be fixed a priori based on established thresholds, placed at quantiles to ensure sufficient observations per segment, or estimated directly using data-driven segmented regression algorithms. Ideally, knots should correspond to meaningful real-world cutoffs, such as age groups (<40, 40–65, ≥65 years), clinical thresholds (e.g., systolic blood pressure ≥140 mmHg), laboratory criteria (HbA1c ≥6.5% for diabetes diagnosis), program milestones (pre- vs. post-treatment initiation), or time since diagnosis (0–6 months vs. >6 months).

When determining the number of knots, there is no single "correct" choice; it is fundamentally a model selection problem. Several common strategies are suggested:[2,20–22]

- Theory-driven selection (preferred when possible). Knots are chosen based on established clinical or biological knowledge. For example, placing two knots at eGFR 60 and 30 mL/min/1.73 m^2, corresponding to moderate and severe chronic kidney disease, allowing comparison of risk across stages of kidney impairment.
- Visual inspection (exploratory): Scatterplots with smoothers (e.g., LOESS or splines) can help identify visible slope changes or plateaus. For instance, plotting blood pressure against age may reveal where risk begins to increase more sharply. This method is intuitive and useful for hypothesis generation.
- Model comparison using information criteria (AIC/BIC). Multiple models with different numbers or placements of knots are fitted and compared using AIC or BIC, which balance model fit and complexity.
- Cross-validation. Models with alternative knot structures are compared using out-of-sample prediction error, for example when predicting hospital readmission risk based on length of stay.
- Rule-of-thumb constraints. The number of knots may be limited by sample size to ensure stable estimation (e.g., at least 20–30 observations per segment).

Importantly, clinical relevance should take precedence over purely statistical optimization. Fewer knots enhance interpretability, and their placement should be clearly justified. Sensitivity analyses are essential to demonstrate robustness and confirm that the model predicts well without introducing unnecessary complexity.[2,5,6]

Conclusion

In summary, piecewise regression offers a practical and interpretable approach for modeling nonlinear trends by fitting distinct linear relationships across meaningful data segments. It is particularly useful when detecting or testing a changepoint is a primary objective and when theory or prior evidence supports structural shifts or threshold effects.[4] The method is especially appropriate for longitudinal and intervention studies, where transition periods and slope changes are substantively important.[23] However, piecewise models should not be used without clear evidence of structural change, adequate sample size, or strong justification for introducing breakpoints. Ultimately, when slopes shift, models must adapt—through carefully justified knots and well-defined segments that balance flexibility with interpretability.

Acknowledgements

An artificial intelligence tool, ChatGPT (OpenAI, 2025), was used to provide programming algorithms for generating simulated datasets used in the statistical analyses presented as illustrative examples in the figures.[24] It also offered language support during the preparation of this manuscript. All generated content was carefully reviewed for accuracy and appropriateness. The author revised and edited the material and assumes full responsibility for the final content.

References

1. Pilgrim C. Piecewise-regression (aka segmented regression) in Python. Journal of Open Source Software. 2021:6(68);3859. doi:10.21105/joss.03859.
2. Kappel S. Piecewise regression: when one line simply isn't enough [Internet]. New York: Datadog; 2017 Jul 11 [cited 2026 Feb 5]. <https://www.datadoghq.com/blog/engineering/piecewise-regression/>
3. Pokiya K. Piecewise linear regression [Internet]. Berlin: ResearchGate; 2022 Jan [cited: 5 Feb 2026]. 9 p. <https://www.researchgate.net/publication/357932259_Piecewise_Linear_Regression>
4. Kaizer A. Segmented linear regression (piecewise, changepoint, broken-line regression) [Internet]. Aurora (CO): Alex Kaizer; [cited 2026 Feb 5 Feb]. <https://www.alexkaizer.com/bios_6618/files/bios6618/W16/segmented_regression.pdf>
5. Gkioulekas I, Papageorgiou LG. Piecewise regression through the Akaike information criterion using mathematical programming. IFAC-Papers OnLine. 2018:51(15);730–5.
6. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol. 2017 Feb 1;46(1):348–55. doi:10.1093/ije/dyw098. Erratum in: Int J Epidemiol. 2020 Aug 1;49(4):1414. doi:10.1093/ije/dyaa118.

7. Lu K-P, Chang S-T. An advanced segmentation approach to piecewise regression models. Mathematics. 2023;11(24):4959. doi:10.3390/math11244959.

8. ∫dw. Piecewise polynomial regression -- spline series (1/4) [Internet]. San Francisco (CA): Medium; 2024 Feb 17 [cited 2026 Feb 5]. <https://medium.com/@JiaHsi/piecewise-polynomial-regression-spline-series-1-4-cb5814707a29>

9. Wicklin R. Piecewise regression models and spline effects [Internet]. Cary (NC): SAS; 2017 Apr 5 [cited 2026 Feb 5]. <https://blogs.sas.com/content/iml/2017/04/05/nonsmooth-models-spline-effects.html>

10. Huang Q. Hands-on tutorial for piecewise linear mixed-effects models using SAS® PROC MIXED. Tallahassee (FL): LexJansen.com; [cited 2026 Feb 5]. <https://www.lexjansen.com/pharmasug-cn/2015/ST/PharmaSUG-China-2015-ST08.pdf>. Conference Proceedings: Pharmaceutical SAS Users Group China 2015; 2015 Sep 4-5; Shanghai, China.

11. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. 6th ed. Hoboken (NJ): Wiley; 2021 Feb. 704 p.

12. Li Y, Li M, Zhang L. Evolutionary polynomial regression improved by regularization methods. PLoS One. 2023 Feb 17;18(2): e0282029. doi:10.1371/journal.pone.0282029.

13. Apurbalal S, Soumen M, Arunendu M. Piecewise linear regression: a new approach to predict COVID-19 spreading. IOP Conf Ser Mater Sci Eng. 2021:1020(1):012017. doi:10.1088/1757-899X/1020/1/012017.

14. Draper NR, Smith H. Applied regression analysis. 3rd ed. Hoboken (NJ): Wiley; 1998 Apr 9. doi:10.1002/9781118625590.

15. Germano MF. Piecewise linear regression model. What is it and when can we use it? [Interent]. San Francisco (CA): Medium; 2021 Jan 1 [cited 2026 Feb 5]. <https://medium.com/data-science/piecewise-linear-regression-model-what-is-it-and-when-can-we-use-it-93286cfee452>

16. Simon L, Young D. Stat 501: Regression Methods. 8.8 Piecewise Linear Regression Models [Interntet]. University Park (PA): PennState, Eberly College of Science; [cited 2026 Feb 5]. <https://online.stat.psu.edu/stat501/lesson/8/8.8>

17. Verboon P, Duif M, Tuijl P. Single Case Design Analyses [Internet]. San Francisco (CA): GitLab Inc; 2024 Oct 15. Chapter 3: Piecewise regression. [cited 2026 Feb 5]; p 23–30. <https://ou-books.gitlab.io/scda---single-case-design-analyses/pwr.html>

18. Wilbert J, Bornert-Ringleb M, Luke T. Statistical power of piecewise regression analyses of single-case experimental studies addressing behavior problems. Front Educ. 202;7:917944. doi:10.3389/feduc.2022.917944

19. Vittinghoff E, Shiboski SC, Glidden DV, McCulloch CE. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. New York: Springer; 2012. doi:10.1007/978-1-4614-1353-0.

20. Muggeo VMR. Segmented: an R package to fit regression models with broken-line relationships. R News. 2008 May;8(1):20–25.

21. Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied linear statistical models. 5th ed. Boston: McGraw-Hill; 2005.

22. Harrell FE. Regression modeling strategies. 2nd ed. New York: Springer; 2015. doi:10.1007/978-3-319-19425-7.

23. Zhang X, Wu K, Pan Y, Yin R, Zhang Y, Kong D, et al. Optimized segmented regression models for the transition period of intervention effects. Glob Health Res Policy. 2023 Jul 24;8(1):29. doi:10.1186/s41256-023-00312-3.

24. OpenAI. ChatGPT, May 6 version [Internet]. San Francisco (CA): OpenAI; 2025 [cited 2026 Feb 5]. <https://chat.openai.com>