مجلة الغري للعلوم الاقتصادية والادارية

مجلد (21) عدد (4) 2025

تقدير المعلمات المقارنه في نمذجة الطقس باستخدام طريقتي الانحدار كينك والانحدار الخطي العادي مع المقادر المربعات الصغرى

شاد توفيق محمد ١
جامعة السليمانية ، كلية الإدارة والاقتصاد
shad.muhamad@univsul.edu.iq

ا.م.د اخترخان صابر حمد 2
جامعة السليمانية ، كلية الإدارة والاقتصاد
akhterkhan.hamad@univsul.edu.iq

المستخلص

تعد تقلبات درجات الحرارة مؤشراً رئيسياً على تغير الأنماط الجوية ولها تأثير مباشر على التخطيط البيئي والتوعية العامة. في هذه الدراسة تم مقارنة بين نموذج الانحدار الخطي المتعدد التقليدي ونموذج الانحدار المنحني المعزز kink ، الذي تم تصميمه لأخذ التغير الهيكلي في تأثير الرطوبة عند حد معين في الاعتبار واخذت بيانات الطقس المجمعة من السليمانية - العراق خلال الفترة من ٢٠٠٥ إلى ٢٠٢٤ لتحليل كيفية تأثير الرطوبة، البخار، واتجاه الرياح على اتجاهات درجات الحرارة. بدأت الدراسة بتطبيق نموذج الانحدار الخطي العادي لنمذجة درجة الحرارة كمتغير تابع. وبسبب التغيرات الملحوظة في العلاقة بين الرطوبة ودرجة الحرارة، تم إدخال نموذج الانحدار المنحني لالتقاط هذا التحول بشكل أكثر فعالية. الانحدار الكينكي هو أسلوب إحصائي يُستخدم لتقدير النماذج التي يتغيّر فيها ميل العلاقة بين المتغيّرات عند نقطة عتبة معيّنة (معروفة أو غير معروفة)، بحيث يُحدث ما يُعرف بـ 'الانكسار' في الميل دون انقطاع في العلاقة. علاوة على ذلك، أظهر هذا النهج المنحني، الذي يستخدم الانحدار الخطي المقطعي، تحسنًا في أداء النموذج. على وجه الخصوص، وذلك بتقليل قيمه المتوسط الخطأ التربيعي (MSE)، وزيادت قيم معامل التحديد (R^2)و معامل التحديد المعدل(Adjusted R^2) ، مما يدل على تحسين التوافق مع البيانات وزيادة موثوقية التنبؤات. حيث يبين هذه النتائج على فعالية استخدام النماذج المعتمدة على الحدود في الدراسات البيئية وتبرز القيمة المضافة لنموذج الانحدار المنحني في الكشف عن العلاقات الهيكلية الخفية ضمن البيانات المناخية.

الكلمات المفتاحية: *الانحدار المتعدد، نموذج الانحدار كينك، تقدير الانحدار الخطي العادي، الانحدار المقطعي*

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (4) 2025

Comparative Parameter Estimation in Weather Modeling Using Kink Regression and Ordinary Least Squares Methods

Shad Tofiq Muhamad
University of Sulaimani, College of Administration and Economics
shad.muhamad@univsul.edu.iq

Assistant Prof Dr. Akhterkhan Sabr hamd
University of Sulaimani, College of Administration and Economics
akhterkhan.hamad@univsul.edu.iq

Abstract

Temperature fluctuations are a key indicator of changing weather patterns and have a direct impact on environmental planning and public awareness. This study utilizes historical weather data collected from Sulaymani, Iraq, covering the period from 2005 to 2024, to analyze how humidity, vapor, and wind direction influence temperature trends. Additionally, the analysis is based on a comparison between the traditional multiple linear regression model and an enhanced kink regression model, designed to account for a structural shift in the effect of humidity at a specific threshold. The study initially applies Ordinary Least Squares multiple regression to model temperature as the dependent variable. However, due to observable changes in the relationship between humidity and temperature, a kink model is introduced to capture this shift more effectively. Kink regression is a statistical method used to estimate models where the relationship between variables changes slope at a known or unknown threshold point, creating a "kink" rather than a discontinuity. Moreover, this kink approach, which utilizes piecewise linear regression, revealed improved model performance. Specifically, the mean square error (MSE) decreased, and both the R^2 and adjusted R^2 increased, indicating a better fit and more reliable predictions. These findings validate the use of threshold-based models in environmental studies and highlight the added value of kink regression in uncovering hidden structural relationships within climatic data.

Keywords: *Multiple Regression, King Regression Model, OLS Estimation, Piecewise Regression.*

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

Introduction

Regression is the statistical technique used for estimating a dependent variable (response) from more than one independent variables (predictors). Additionally, the crucial purpose of regression analysis is prediction and causal inference. Estimating the dependent variable with a given set of independent variables is known as prediction. While causal inference explains the relationship between a dependent variable and one or more independent variables. Regression analysis is applicable across a wide range of disciplines, including econometrics, healthcare, marketing, finance, engineering, weather …etc. [10] [19] [24]

Linear function or parametric functional form is one of the most traditional regression models, which applies continuous relationships among variables. Linear Regression model is one of the most fundamental and widely used in statistical methods for estimating the relationship between a single dependent variable and one or more independent variables that observed data can be fitted by a linear equation line. It indicates that a single linear line represents the best fit through the data points, explaining how changes in independent variables lead to change in the dependent variable. [21] [25]

Regression assumptions is crucial step in the regression analysis that ensure the validity and relaibility of the model results. These assumptions like linearity, independence of errors, homoscedasticity, normality of errors, and no multicollinearity. Having these assumptions lead to get an accurate coefficient estimates and valid statistical inferences. Other than that, vice versa. [13] [17]

While real world experiences expand the need for data, collecting a vast amount of data is challenging, especially when trying to apply all

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

assumptions without violating them. That is where complex model of regression can face these challenges such as non-linear, quasi-experiment, and multiple regression etc... In the case, if linear regression fails to capture the true pattern model. This is where quasi-experiment provides an alternative method for handling approach such as, Regression Discontinuity Design or Regression Kink Design. [4] [18]

Regression Kink Design is a quasi-experiment used to indicate and identify the relationship between one dependent variable and one of more independent variables based on on a cetrain point which is known as threshold. Regression Kink Design is similar to Regression Discontinuity Design. Regression Discontinuity Design relies on abrubt jumps in the outcome variable to identify casual effects at the threshold or cutoff. Meanwhile, Regression Kink Design apply a kink point at the threshold rather than a jump. [2] [6] [7]

Material and Methods:

2.1 Regression model: [31] [10] [28] [30]

2.1.1 Simple Linear Regression

Simple linear regression is one of the simplest statistical methods used to determine the effect of one predictor on outcome, these two variables assume have a linear correlation line. The line is determined by minimizing the sum of squared errors between the observed values and the predicted values. [8] [12] [19]

$$(1) \boldsymbol{Y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i}$$

Y_i: is the outcome variable.

β_0: is the intercept (the value of Y when is x_1 is 0).

β_1: is the slope (the change in Y for a one-unit increase in x_1).

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

x_1: is the independent variable (the predictor).

ε_i: it represents the error term i^{th} observations.

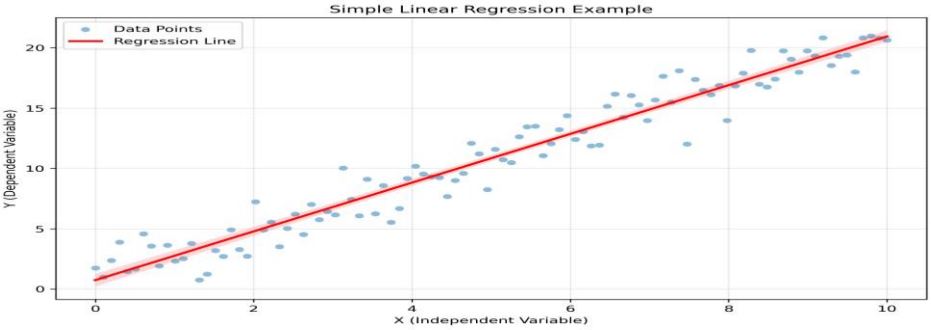

Figure 1. This graph explains simple linear regression between one dependent variable and one independent variable.

2.1.1 Multiple Linear Regression

This regression is an extension form of simple linear regression that is used to identify and estimate and determine the value of a dependent variable based on one or multiple independent variables. Furthermore, it explains how multiple independent variables collectively effect on a dependent variable. Unlike simple regression. Each independent variable's coefficient shows its impact while controlling for others. This method helps indicate more complex situations where multiple factors influence one outcome. [1] [19] [27]

$$\boldsymbol{Y_i = \beta_0 + \beta_1 x_1 + \ \beta_2 x_2 + \beta_n x_m + \varepsilon_i} \qquad \textbf{(2)}$$

Where:

Y_i : is the dependent variable (the outcome we are trying to predict) for the i^{th} observation.

x_1, x_2..., x_m : are the independent variables (the predictors) for the i^{th} observation.

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

β_0: is the intercept of the regression line, representing the value of Y_i when all the X's are zero.

$\beta_1, \beta_2 \ldots, \beta_n$: are the coefficients that represent the change in the dependent variable. Each β_i represents the change in Y_i for a one-unit increase in the respective x_i, holding all other variables constant.

ε_i : The error term for the i^{th} observation, which accounts for the variability in Y_i that cannot be explained by the linear relationship with the independent variables.

2.2 Regression Model Assumptions

Regression assumptions are the foundational conditions that verify both reliability and the validity of a regression model's results. These assumptions are key to guaranteeing that estimated coefficients are unbiased, consistent, and efficient. In the meantime, any violation of these assumptions will lead to an incorrect conclusion. Essentially, they establish the basis that guarantees the accuracy and relevance of regression analysis, reducing errors that may distort the relationships among variables: [13] [17] [21]

2.2.1 Linearity

The linearity assumption in regression analysis states that The model assumes a linear relationship, meaning that any variation in the independent variables should lead to proportional changes in the dependent variable. The assumption ensures that the model can accurately capture the relationship between variables using a straight line or hyperplane. Violating this assumption can lead to biased estimates and incorrect conclusions. It is essential to verify linearity before fitting a regression model to ensure reliable results. [13] [22]

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

A scatter plot of residuals can reveal linearity. If the residuals are randomly scattered with no pattern, the relationship is linear. A curved or systematic pattern in the residuals suggests a non-linear relationship.

Hypothesis Test:

H_0: The relationship is linear.

H_1: The relationship is non-linear.

2.2.2 Constant Error Variance (Homoscedasticity)

Homoscedasticity is another assumption that shows the variance of the residuals (errors) is constant across all levels of the independent variables. This means that the spread of the residuals should remain the same throughout the range of fitted values. Violating this assumption (i.e., heteroscedasticity) can lead to inefficient estimates and biased statistical tests. It is important to check for homoscedasticity to ensure the model's reliability. [17] [30]

One of the most common methods to detect heteroscedasticity is Breusch-Pagan test heteroscedasticity by testing whether the variance of the residuals is constant across all levels of the independent variables.

$$\text{LM} = \text{n}.R^2 \qquad (3)$$

LM: Lagrange Multiplier statistic (Breusch-Pagan test statistic).

n: number of observations.

R^2: coefficient of determination from the auxiliary regression of ε_i on the independent variables.

Hypothesis Test:

H_0: Homoscedasticity included into data set.

H_1: Heteroscedasticity included into data set.

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

2.2.3 Independent of Error Term

The independence of error terms assumption in regression analysis indicates that the residuals (errors) of the model should be independent of each other. This means that the value of the error term for one observation should not provide any information about the value of the error term for another observation. In other words, there should be no autocorrelation among the errors. The Durbin-Watson test is used to detect autocorrelation in the residuals of a regression model. [13]

The Durbin-Watson test is used to detect autocorrelation in the residuals of a regression model, specifically focusing on first-order autocorrelation. A value close to 2 suggests no autocorrelation, while values significantly below or above 2 indicate positive or negative autocorrelation, respectively. [1]

$$\text{DW} = \frac{\Sigma_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} \boldsymbol{e}^2 t} \qquad (4)$$

e_t = residual (error term) at time t.

n = number of observations.

<u>Hypothesis Test:</u>

H_0: Autocorrelation has not affected the data.

H_1: Autocorrelation has affected the data.

2.2.4 Normal Error

Normal Error is a crucial property all over assumptions. This assumption explains regression analysis states that the residuals (errors) of the model should be normally distributed. Normality ensures that the statistical significance of the model's parameters can be accurately assessed. [30]

To indicated normality Kolmogorov-Smirnov (K-S) test is one of most usable tests to detect normality.

$$D_n = \sup_x |F_n(x) - F(x)| \qquad (5)$$

D_n: K-S statistic

$F_n(x)$: empirical distribution function (EDF) of the sample

$F(x)$: cumulative distribution function (CDF) of the reference distribution (e.g., normal)

$\sup_x$: supremum (maximum) value of the absolute differences over all x.

Hypothesis Test:

H_0: Errors are normally distributed.

H_1: Errors are not normally distributed.

2.2.5 No Multi-Collinearity

No multicollinearity is a common assumption in multiple regression analysis that the independent variables should not be highly correlated with each other. When independent variables are highly correlated, it becomes difficult to determine the individual effect of each variable on the dependent variable, leading to unstable coefficient estimates and inflated standard errors. This can make statistical tests unreliable, resulting in biased or misleading results. Multicollinearity is often assessed using the Variance Inflation Factor (VIF), where a VIF value greater than 10 typically indicates problematic multicollinearity. [29]

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2} \qquad (6)$$

VIF_j: Variance Inflation Factor for predictor j.

R_j^2The coefficient of determination (R-squared) when the j^{th} predictor is regressed on all the other predictors in the model.

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

Hypothesis Test:

H_0: Multicollinearity is not existed.

H_1: Multicollinearity is existed.

2.3 Regression Kink Design

First appearance of Regression Kink Design was laid at the very late of 2015 in the field of econometric by David S. Lee and Thomas Lemieux. Regression Kink Design is known as quasi-experiment that determine and estimate the casual effect between regression variables that based on a cetrain point which is known as threshold. Regression Kink Design apply a kink point at the specific point where threshold is appear to identify the behaviour of data. [3] [7] [16]

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + \beta_{n+1}(x_{1i} - x^*)D_i + \varepsilon_i \qquad (7)$$

$\boldsymbol{Y_i}$: is the dependent variable for the i^{th}observation.

$\boldsymbol{\beta_0}$: The intercept term.

$\boldsymbol{x_{1i}, x_{2i}, \dots, x_{ni}}$: The independent variables (predictors) for the i^{th} observation

$\boldsymbol{\beta_1, \beta_2, \dots, \beta_n}$: The coefficients for the independent variables, representing the effect of each independent variable on Y_i.

$\boldsymbol{x^*}$: The threshold or cutoff point

$\boldsymbol{D_i}$: A binary indicator variable (dummy variable) where $D_i = 0$ when $x_{1i} \leq x^*$ and $D_i = 1$ when $x_{1i} \geq x^*$. This variable is used to trigger the threshold effect.

$\boldsymbol{\beta_{n+1}}$: The coefficient for the interaction term $(x_{1i} - x^*)D_i$, capturing the change in the relationship between x_1 and Y_i once x_1 exceeds the threshold x^*.

$\boldsymbol{\varepsilon_i}$: The error term, which captures the residuals (unexplained variance).

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

2.4 Regression Kink Assumption

Regression Kink Design relies on some key assumptions to ensure validity causal inference. These assumptions include the continuity of the outcome at the threshold, absence of manipulation in the running variable, and proper model specification. In other hand, violations each one of these assumptions may show bias the estimated kink effect and lead to incorrect conclusions. [6] [14]

2.4.1 Continuity of Potential Outcomes at the Threshold

This assumption continuity at threshold point is key function that requires the outcome variable would follow a smooth trend around the threshold if there were no treatment. Furthermore, it explains that any observed value change in the slope at the threshold is due to the treatment effect, not a jump in the outcome level. [6] [14]

$$\lim_{X\to C^-} E[Y_i(0)|X_i = X] = \lim_{X\to C^+} E[Y_i(0)|X_i = X] \qquad (8)$$

This formula states that the expected outcome without treatment is continuous at the threshold from both sides (left and right). It ensures that any observed change in the slope at the kink point is due to the treatment, not a jump in the outcome level.

2.4.2 No Sorting or Manipulation of the Running Variable

No Manipulation of the Running Variable is a crucial assumption in all other. Eventually, means individuals cannot precisely control the running variable (e.g., income) to position themselves around the threshold. If manipulation exists, it may bias the estimation by introducing selection effects. [6] [14]

$$\lim_{x\to C^-} f_x(x) = \lim_{x\to C^+} f_x(x) \qquad (9)$$

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

This formula means the distribution of the running variable should be smooth at the threshold — no sudden jump in values. It ensures there's no manipulation or bunching of observations just above or below the kink point.

2.5 Piecewise Regression Model

The piecewise regression model is a statistical approach used to identify shifts in relationships that occur at certain predefined points within the data that often referred to "Cut-Off" or "Thresholds. Piecewise regression provides more accurate representation of non-linear relationship that exhibits different trends or behaviors before and after a specific threshold. This approach offers a better fit by accounting for changes in the relationship among variables. Unlike standard linear regression, which assumes a single linear relationship through all data, piecewise regression divides the data into segments and fits separate linear models to each segment. This method is particularly useful when the relationship between the independent and dependent variables is not constant but changes at specific values of the independent variable which is called threshold. By introducing these threshold, piecewise regression provides a more accurate representation of data with varying trends, such as in economics, environmental studies, and other fields where the relationship shifts across different ranges. [20] [26]

2.6 Ordinary Least Square Estimation

In parametric statistics, the researcher utilizes a sample statistic as an approximation of the population parameter. Estimate is the value that taken from the sample size which explains the population's true value. Moreover, estimation is the value of a sample statistic that provides information about the population parameter. There are some important properties of estimations that leads the sample statistic to be an accurate and stable

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

estimate of the population parameter. Unbiased, consistent, efficient, and sufficient are the most important of behavior of each estimate. Additionally, If the data satisfies all the regression assumptions, it implies that the data is unbiased, consistent, efficient, and sufficient. This allows us to apply one of the most reliable estimation methods. [5] [15]

Ordinary Least Squares is a statistical technique for estimating the coefficients of a linear regression model by minimizing the total of the squared differences between observed and predicted values, which are the differences between observed and predicted values. Ordinary Least Squares assumes that the errors are independently and identically distributed with a mean of zero and constant variance. It provides the best linear unbiased estimators (BLUE) under the Gauss-Markov theorem, ensuring that the estimates are unbiased and have the smallest possible variance among all linear estimators. Ordinary Least Squares is optimal when the model meets its underlying assumptions.

$$\hat{B} = (x^T x)^{-1} x^T y \qquad (10)$$

In OLS regression, $\hat{B}$ represents the vector of estimated coefficients, which includes both the intercept and the slope. X is the matrix of independent variables, with a column of 1s added for the intercept, while y is the vector of observed dependent variable values.

2.7 Analysis Of Variance (ANOVA)

ANOVA is a statistical technique used to compare the amount of variation explained by a model to the variation left unexplained. Additionally, it tests whether different variables in a model explain significant variation in the outcome. It evaluates whether the model's predictors improve the fit significantly more than chance, using F-statistics. Moreover, the total variance is split into model sum of squares and error sum of squares. Mean Squared Error (MSE) is the average of the error sum of squares per degree of freedom and reflects the variance not explained by the model. MSE explains how much error, on average, remains unexplained by the model. [9] [23]

2.8 The Wald Test

The Wald test is a statistical procedure used to determine whether a notable change occurs in the relationship between two variables at a particular point, known as the threshold. In kink regression, it helps determine if the slope of the relationship changes before and after the threshold. The test compares the slopes on either side of the threshold and checks if they are different. A significant result means the relationship between the variables changes at the threshold. This helps identify important shifts or changes in patterns within the data. [16]

$$\text{W} = \frac{(\hat{B}-B_o)^2}{\text{var}(\hat{B})} \sim x_1^2 \qquad (11)$$

The Wald test statistic, W, compares the estimated coefficient $\hat{B}$ to a hypothesized value B_o under the null hypothesis. And Walt follows chi-squared distribution with 1 degree of freedom.

<u>Hypothesis Test:</u>

***H*₀:** There is no kink (slopes are equal on both sides).

***H*₁:** There is a kink (slopes differ).

Result and Discussions

3.1 Describe of data:

In this research, weather data from Sulaymaniyah spanning from 2005 to 2024 has been analyzed, focusing on key environmental factors. The data was obtained from reliable sources which include very detailed data. The

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

study focuses on examining the effect of 3 factors that affect temperature. Moreover, various statistical methods such as python and R programming can be used to analyze this data. The temperature was used as the response variable y, while humidity x_1, vapor x_2, and wind direction x_3 were included as explanatory variables. This data provided valuable insights into how these factors have influenced temperature trends over the given period.

3.2 Appling Multiple Linear Regression

$\hat{y}_i = 1.492662089e^{-15} - 0.6947\hat{x}_1 + 0.3656\hat{x}_2 + 0.0401\hat{x}_3$ depending on equation (2)

The equation explains that, the temperature decreases by 0.6947 units for every single unit increase in humidity, and increases by 0.3656 units for every single unit increase in vapor, lastly increases by 0.0401 units for every single unit increase in wind direction.

3.3 Testing All Regression Assumption

3.3.1 Linearity

The Residuals plot is one of the most common ways to identify whether the data is following linearity or not. If the residuals appear to be randomly distributed around zero without any obvious pattern, it indicates that the assumption of linearity is likely valid the assumption holds true. In other hand, if the plot shows any systematic pattern or curvature, it indicates this may suggest that the connection between the dependent and independent variables is potentially non-linear.

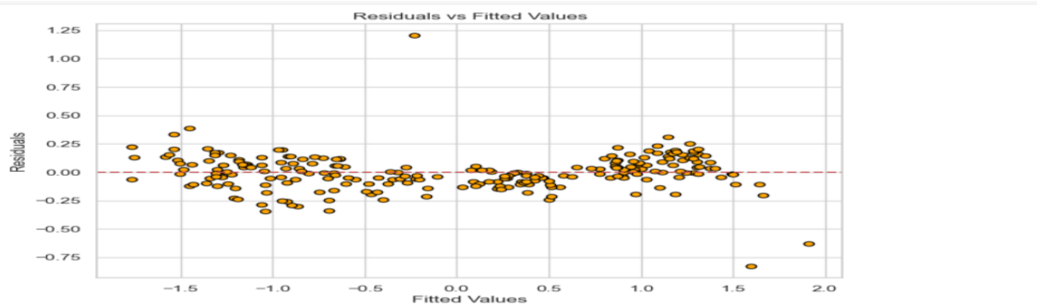

Figure 2. This graph does not appear any clear pattern or curve, and mostly around zero, the data indicates linearity.

3.3.2 All Other Assumptions:

Table 1. This test shows all regression assumptions result and detecting

Test	Detection	Result	P-Value
Breusch-Pagan	Constant Error Variance	constant variance holds	0.7454
Durbin-Watson	Independent of Error Term	No Autocorrelation	1.8145
Kolmogorov-Smirnov	Normal Error	The data is normal	0.1129
VIF for X1	Multi-Collinearity	No Multi-Collinearity	1.771024
VIF for X2		No Multi-Collinearity	1.743538
VIF for X3		No Multi-Collinearity	1.316737

The table present key diagnostic tests for examining the assumptions of multiple linear regression are included in Table. The Breusch-Pagan test has a p-value of 0.7454, which means that homoscedasticity is also satisfied since there is no heteroscedasticity in the residuals. The Durbin- Watson statistic of 1.8145 indicates no autocorrelation in the residuals, validating the assumptions of the errors independence. Also, the p-value of the Kolmogorov-Smirnov test for normality is 0.1129, thus also the residuals are normally distributed. The Variance Inflation Factor (VIF) values for the independent variables (x_1, x_2, and x_3) are, likewise, 1.77, 1.74, and 1.31, respectively—all less than the typical cut-off value of 10—suggesting no multicolinearity. Combined, these diagnostics demonstrate that the regression model meets the fundamental statistical assumptions necessary for the accuracy of the model's estimates and inferences.

Table 2. This Table Explains OLS Estimation before applying kink model.

	R^2	Adj. R^2	F-Test	Log-Likelihood
OLS	0.973	0.973	2812	92.991

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

This table provides the OLS model's performance before kink regression. The model has demonstrated good fit with an R2 and adjusted R2 equal to 0.973, a very high F-statistic of 2812 which means joint significance and a log-likelihood of 92.991 indicating good model adequacy.

3.4 Stepwise Regression

Stepwise regression is a statistical approach used to identify and retain the most important variables within a regression model. Furthermore, by adding or removing predictors based on specific criteria, such as the p-value or AIC (Akaike Information Criterion).

Table 3. This Table Explains Stepwise regression to detect if the variables are having effect on the morel or not

	coefficient	standard error	t	P>|t|	0.025	0.975
Const	5.239e-16	0.011	4.89e-14	1.000	-0.021	0.021
Humidity	-0.6947	0.014	-48.760	0.000	-0.723	-0.667
Vapor	0.3656	0.014	25.942	0.000	0.338	0.393
Wind Direction	0.0401	0.012	3.291	0.001	0.016	0.064

In the stepwise regression output, Avg Humidity, Avg Vapor, and Wind Direction are statistically significant with p-values less than 0.05, indicating they affect the model. The constant term is not significant (p-value = 1.000) and likely doesn't contribute meaningfully to the model.

3.5 Analysis Of Variance (ANOVA)

Analysis of Variance is a statistical technique used to compare the variation explained by a model with the unexplained variation. It is used to assess whether different variables in a model significantly influence the outcome. ANOVA helps determine if the predictors improve the model’s fit compared to chance.

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

Table 4. This Table shows ANOVA before applying kink model.

Source Of Variance	Degree of freedom	Sum Of Square	Mean Sum Of Square	F-value	P-value
Humidity	1	64.182	64.182	2377.58	0.000
Vapor	1	18.167	18.167	672.98	0.000
Direction	1	0.292	0.292	10.83	0.001
Regression	3	227.764	75.9214	2812.48	0.000
Residual	231	6.236	0.0270	-	-

This ANOVA is a measure of the relative contribution of the predictors prior to applying the kink model. Then humidity has the greatest contribution to the model, with an F-value of 2377.58, followed by vapor (672.98) and direction (10.83), being all statistically significant ($p < 0.01$). The whole model is highly significant ($F = 2812.48$, $p = 0.000$), which means that the model is strong. The residual sum of squares (6.236) is low, indicating a good model fit with small unexplained variance.

3.6 Regression Kink Assumptions:

Table 5. This table shows the test of kink model assumptions with detections

Test	Detection	Result	P-Value
Wald Test	To Test If There Is Kink	There Is Kink	0.000
Placebo Test	If The Kink Point at threshold is Significant	The Kink Point at threshold is Significant	0.0501

To validate the presence of a structural change in the model, both the Wald and Placebo tests were conducted. Results confirm a statistically significant kink point at the estimated threshold.

3.7 Kink Regression Model:

Kink Regression model is one of the most advance models that recently developed. This model can be used to estimate relationship between variables based on a certain threshold to ensure if the data can sudden change

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

at that threshold and detect the effect of the threshold and show the level of changes. The best way to create a regression model based on a threshold is piecewise regression. Piecewise Regression is also known as segmented regression. This model of regression is designed to fit different linear models to different segments of the data from right and left threshold. Piecewise Regression is preferred to determine the relationships between a dependent and an independent variable based on a change based on a certain point called threshold while all regression assumptions are met.

$\hat{y}_i = -0.2627 - 0.9667\hat{x}_1 + 0.1574\hat{x}_2 + 0.3141\hat{x}_1\hat{x}_2 + 0.3637\hat{x}_3 + 0.3637\hat{x}_3 + 0.0405\hat{x}_4$ based on equation (7) we can estimate parametes. The model estimates that before the humidity threshold of −0.26, temperature drops sharply (−0.97 units per humidity unit). After the threshold, this effect weakens to −0.57, with a small jump in level (+0.16). Vapor significantly increases temperature (+0.36 per unit), and wind direction has a smaller positive effect (+0.04). The intercept (−0.26) reflects baseline temperature when all inputs are zero.

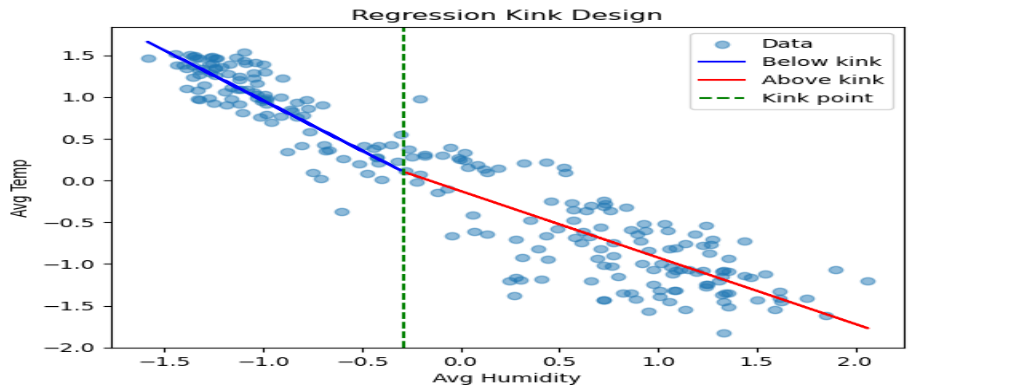

Figure 3. This explains the threshold at -0.29 and separate the left linear line from the right linear line in order to see whether kink can apply or not.

Table 6. This Table Explains piecewise regression to identify linear line from the left side and write side to detect kink.

	coefficient	standard error	t	P>|t|	0.025	0.975
const	-0.2627	0.051	-5.163	0.000	-0.363	-0.162
Humidity	-0.9667	0.049	-19.873	0.000	-1.063	-0.871
Threshold	0.1574	0.056	2.832	0.005	0.048	0.267
Kink Humidity	0.3941	0.053	7.374	0.000	0.289	0.499
Vapor	0.3637s	0.013	28.888	0.000	0.339	0.388

Table 7. This Table detect the OLS estimation after apply kink on the model**.**

	R^2	Adj. R^2	F-Test	Log-Likelihood
OLS	0.979	0.978	2661	120.15

After applying the kink model, all variables remain statistically significant ($p < 0.01$), with "Kink Humidity" showing a strong positive effect (β = 0.3941).

The threshold variable is also significant ($p = 0.005$), confirming a slope change at the kink point. Model fit improved with R^2 = 0.979, Adjusted R^2 = 0, and a higher log-likelihood (120.15), indicating better explanatory power. The F-test value (2661) confirms overall model significance.

Table 8. This Table shows ANOVA after applying kink model

Source Of Variance	Degree of freedom	Sum of Square	Mean Sum of Square	F-value	P-value
Humidity	1	207.569220	207.569220	9640.360143	0.000
Above Threshold	1	0.234408	0.234408	10.886839	0.000
Kink Humidity	1	1.274761	1.274761	59.205094	0.000
Vapor	1	19.693303	19.693303	914.637212	0.000
Wind Direction	1	0.297647	0.297647	13.823950	0.000
Residual	229	4.930661	0.021531	-	-

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

This ANOVA table summarizes the significance of predictors in a kink regression model. Humidity and Vapor exhibit the highest explanatory power (F-values: 9640.36 and 914.64, respectively), with all predictors showing statistically significant effects (P-values = 0). The residual variance is minimal, indicating a strong overall model fit.

4. Conclusion And Recommendations:

4.1 Conclusion:

Relying upon the findings obtained from the practical component of the study, a number of significant conclusions can be derived, as outlined below: the study initially applied multiple linear regression using the OLS method to model the relationship between temperature (dependent variable) and multiple predictors, including humidity. A kink regression model was later introduced by applying a threshold to humidity, allowing for a piecewise linear structure. The results showed a notable improvement in model performance: Mean Squared Error (MSE) decreased, while both R^2 and Adjusted R^2 increased. This indicates that the kink model better captured the linear relationship between humidity and temperature, validating the decision to introduce a structural break in the predictor. Finaly, the results of the kink regression revealed a slope of −1.2114 before the threshold (Humidity ≤ −0.29) and −0.8004 after the threshold (Humidity > −0.29). The treatment effect, represented by the change in slope at the threshold, was 0.4110. This shift suggests that while the relationship remains linear, the rate at which temperature responds to changes in humidity differs across humidity levels — a nuance effectively captured by the kink model.

4.2 Recommendations:

- **Use Different Threshold and Kink Points**

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

Try applying the kink model with different humidity threshold values instead of fixing it at −0.29. This can help find a more accurate or meaningful turning point in the relationship.

- **Apply Multiple Kink Points**

To better capture complex nonlinear behavior, a model with multiple kink points (piecewise linear segments) should be considered. This approach can identify more than one significant slope change, which may occur due to environmental thresholds or interactions between predictors.

- **Use Real Values and Robust Estimation**

Future models should utilize the original (non-standardized) units of the predictors to enhance interpretability and practical application. Additionally, incorporating robust regression techniques (e.g., Huber or M-estimators) can reduce the influence of outliers and improve model reliability in real-world data conditions.

5. References:

1. Aiken, L.S., West, S.G., Pitts, S.C., Baraldi, A.N. and Wurpts, I.C., 2012. Multiple linear regression. Handbook of Psychology, Second Edition, 2.
2. Ando, M., 2017. How much should we trust regression-kink-design estimates ?. Empirical Economics, 53, pp.1287-1322.
3. Bana, S.H., Bedard, K. and Rossin-Slater, M., 2020. The impacts of paid family leave benefits: regression kink evidence from California administrative data. Journal of Policy Analysis and Management, 39(4), pp.888-929.
4. Bärnighausen, T., Tugwell, P., Røttingen, J.A., Shemilt, I., Rockers, P., Geldsetzer, P., Lavis, J., Grimshaw, J., Daniels, K., Brown, A. and Bor, J.,

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

2017. Quasi-experimental study designs series—paper 4: uses and value. Journal of clinical epidemiology, 89, pp.21-29.

5. Burton, A.L., 2021. OLS (Linear) regression. The encyclopedia of research methods in criminology and Criminal Justice, 2, pp.509-514.
6. Card, D., Lee, D.S., Pei, Z. and Weber, A., 2015. Inference on causal effects in a generalized regression kink design. Econometrica, 83(6), pp.2453-2483.
7. Card, D., Lee, D.S., Pei, Z. and Weber, A., 2017. Regression kink design: Theory and practice. In Regression discontinuity designs: Theory and applications (pp. 341-382). Emerald Publishing Limited.
8. Chatterjee, S., & Hadi, A. S. (2012). Regression Analysis by Example (5th ed.). Wiley.
9. Cuevas, A., Febrero, M. and Fraiman, R., 2004. An anova test for functional data. Computational statistics & data analysis, 47(1), pp.111-122.
10. Diskin, M.H., 1970. Definition and uses of the linear regression model. Water Resources Research, 6(6), pp.1668-1673.
11. Dong, Y., 2016. Jump or kink? Regression probability jump and kink design for treatment effect evaluation. Unpublished manuscript.
12. Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (3rd ed.). Wiley.
13. Flatt, C. and Jacobs, R.L., 2019. Principle assumptions of regression analysis: Testing, techniques, and statistical reporting of imperfect data sets. Advances in Developing Human Resources, 21(4), pp.484-502.
14. Ganong, P. and Jäger, S., 2018. A permutation test for the regression kink design. Journal of the American Statistical Association, 113(522), pp.494-504.
15. Gaure, S., 2013. OLS with multiple high dimensional category variables. *Computational Statistics & Data Analysis*, *66*, pp.8-18.
16. Gregory, A.W. and Veall, M.R., 1985. Formulating Wald tests of nonlinear restrictions. Econometrica: Journal of the Econometric Society, pp.1465-1468.

17. Hayes, A.F. and Cai, L., 2007. Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. Behavior research methods, 39, pp.709-722.
18. Maciejewski, M.L., 2020. Quasi-experimental design. Biostatistics & Epidemiology, 4(1), pp.38-47.
19. Marill, K.A., 2004. Advanced statistics: linear regression, part II: multiple linear regression. Academic emergency medicine, 11(1), pp.94-102.
20. McZgee, V.E. and Carleton, W.T., 1970. Piecewise regression. Journal of the American Statistical Association, 65(331), pp.1109-1124.
21. Neter, J., Wasserman, W. and Kutner, M.H., 1989. Applied linear regression models.
22. Poole, M.A. and O'Farrell, P.N., 1971. The assumptions of the linear regression model. Transactions of the Institute of British Geographers, pp.145-158.
23. St, L. and Wold, S., 1989. Analysis of variance (ANOVA). Chemometrics and intelligent laboratory systems, 6(4), pp.259-272.
24. Sykes, A.O., 1993. An introduction to regression analysis.
25. Ter Braak, C.J.F. and Looman, C.W.N., 1995. Regression. In Data analysis in community and landscape ecology (pp. 29-77). Cambridge University Press.
26. Toms, J.D. and Lesperance, M.L., 2003. Piecewise regression: a tool for identifying ecological thresholds. Ecology, 84(8), pp.2034-2041.
27. Tranmer, M. and Elliot, M., 2008. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), pp.1-5.
28. Tranmer, M. and Elliot, M., 2008. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), pp.1-5.
29. Tsagris, M. and Pandis, N., 2021. Multicollinearity. American journal of orthodontics and dentofacial orthopedics, 159(5), pp.695-696.
30. Uyanık, G.K. and Güler, N., 2013. A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences, 106, pp.234-240.

DOI: https://doi.org/10.36325/ghjec.v21i4.19651

31. Wooldridge, J. M. (2016). Introductory Econometrics: A Modern Approach (6th ed.). Cengage Learning.

32. Paravision Lab. (n.d.). Simplify data analysis with simple linear regression in R [Image]. Retrieved May 8, 2025, from https://www.paravisionlab.com/.

DOI: https://doi.org/10.36325/ghjec.v21i4.19651