Goethe-Gymnasium
Schulstraße 36
56130 Bad Ems

Facharbeit

im

Leistungskurs Mathematik

Klassenstufe MSS 12

2022/2023

Betreuender Lehrer: Herr Kreutz

Einfache Lineare Regression mit der Methode «Linear Least Squares»

Beispielanwendung auf ein schulisches Chemieexperiment

Adnan Aidak
Grabenstr. 20
56130 Bad Ems

Abgabe Termin: 12.06.2023

Abstract

Diese Facharbeit beschäftigt sich mit der Anwendung der Linearen Least Squares Methode zur Berechnung einer einfachen linearen Regression. Das Ziel dieser Arbeit ist es, den linearen Zusammenhang zwischen einer abhängigen und einer unabhängigen Variablen zu analysieren und ein Programm zu entwickeln, das die Lineare Least Squares Methode implementiert. Der theoretische Teil der Arbeit befasst sich zunächst mit den mathematischen Grundlagen der einfachen linearen Regression sowie der Least Squares Methode. Anschließend wird die Lineare Least Squares Methode eingehend erläutert. Die Fehleranalyse und Validierung der Methode werden anhand von Residuenanalyse, Ausreißern, dem Korrelationskoeffizienten r und dem Bestimmtheitsmaß R^2 untersucht. Zudem werden alternative Methoden vorgestellt. Im praktischen Teil der Arbeit wird diese Methode auf ein chemisches Experiment angewendet und dafür dem Programm verwendet. Es werden die Schritte zur Datenbeschaffung und Vorverarbeitung erläutert, gefolgt von der Anwendung der Methode auf die Daten und der Analyse der Ergebnisse. Die Interpretation der Ergebnisse bildet einen wichtigen Aspekt der Anwendungen der Linearen Least Squares Methode. Abschließend wird ein Fazit gezogen, in dem die Ergebnisse der Arbeit zusammengefasst und diskutiert werden. Die Facharbeit liefert somit einen umfassenden Überblick über die Anwendung der Linearen Least Squares Methode und eine Anwendung auf ein schulisches Experiment und präsentiert ein entwickeltes Programm zur praktischen Umsetzung der Methode.

Inhalt

1 Einleitung 1

2 Mathematische Grundlagen 2

2.1 Einfache Lineare Regression 2

2.2 Least Squares Methode 2

2.3 Lineare Least Squares Methode 3

3 Fehleranalyse, Validierung und Alternativen der LLS-Methode 5

3.1 Residuenanalyse 6

3.1.1 Ausreißer 6

3.2 Korrelationskoeffizient r 7

3.3 Bestimmtheitsmaß R^2 7

3.4 Andere Methoden 8

4 Implementierung von Linear Least Squares 8

4.1 Programm zur Berechnung von Linear Least Squares 8

4.1.1 Bibliotheken 9

5 Anwendungen der Linearen Least Squares Methode 10

5.1 Datenbeschaffung und -vor Verarbeitung 10

5.2 Einfache Lineare Regression und Analyse 11

5.3 Interpretation der Ergebnisse 12

6 Fazit 12

7 Anhang 14

7.1 Literaturverzeichnis 14

7.2 Code 15

1 Einleitung

Die einfache lineare Regression mit der Methode "Linear Least Squares" ist ein zentrales Thema in der statistischen Analyse und findet in vielen Bereichen Anwendung. Insbesondere in den Naturwissenschaften, wie der Chemie, Physik und Mathematik, ist es oft notwendig, die lineare Abhängigkeit zwischen Variablen zu überprüfen, um Experimente und Phänomene zu verstehen. Jedoch wird dieses Thema im schulischen Unterricht oft nur oberflächlich behandelt und provisorisch angegangen. Aus diesem Grund wurde entschieden, die Facharbeit genau diesem Thema zu widmen. Die einfache lineare Regression bietet eine statistische Methode, um den linearen Zusammenhang zwischen einer abhängigen und einer unabhängigen Variablen zu analysieren. Sie ermöglicht es uns, quantitative Aussagen über die Steigung und den y-Achsenabschnitt einer Regressionsgeraden zu treffen, die die Daten am besten beschreibt. Diese Informationen sind von entscheidender Bedeutung, um Experimente zu planen, Zusammenhänge zu verstehen und Vorhersagen zu treffen. In vielen chemischen und physikalischen Experimenten ist es unerlässlich, die lineare Abhängigkeit zwischen den untersuchten Variablen zu prüfen. Beispielsweise kann die Konzentration eines Stoffes in Abhängigkeit von der Zeit gemessen werden, um den Zerfall oder die Reaktionskinetik zu untersuchen. Oder es kann die Temperatur in Abhängigkeit von der Druckänderung gemessen werden, um die gasförmigen Zustandsänderungen zu analysieren. In solchen Fällen ist es wichtig, die Methode der einfachen linearen Regression anzuwenden, um den linearen Zusammenhang zu bestätigen und fundierte Schlussfolgerungen ziehen zu können. Die Schülerinnen und Schüler erhalten meist nur eine grundlegende Einführung in die Konzepte und Techniken, was zu einem unvollständigen Verständnis und Anwendung führen kann. Daher ist es wichtig, dieses Thema genauer zu erforschen und ein tieferes Verständnis für die Methoden und deren Anwendungen zu entwickeln. Das Ziel dieser Facharbeit ist es, einen umfassenden Überblick über die Methode der einfachen linearen Regression mit der "Linear Least Squares"-Methode zu geben und deren Anwendung in den Naturwissenschaften anhand eines Chemieexperimentes, der die Abhängigkeit der Reaktionsgeschwindigkeit von der Konzentration von Thiosulfat untersucht. Dafür würde ein Programm entwickelt, dass die Darstellung und die Berechnung vereinfacht. Durch eine gründliche Analyse und Darstellung dieser Methodik trägt diese Facharbeit dazu bei, das Verständnis und die Anwendung der einfachen linearen Regression zu verbessern und ihren Wert als statistisches Werkzeug in den Naturwissenschaften zu verdeutlichen.

2 Mathematische Grundlagen

2.1 Einfache Lineare Regression

Die lineare Regression ist eine statistische Methode zur Modellierung des Zusammenhangs zwischen einer abhängigen und einer unabhängigen Variablen. In der einfachen linearen Regression wird der Zusammenhang zwischen einer abhängigen Variable y und einer unabhängigen Variable x betrachtet. Das Ziel besteht darin, eine Ausgleichgerade zu finden, die den Zusammenhang zwischen x und y am besten beschreibt. Die Gerade wird in der Form $y = b + ax$ angenommen, wobei b den y-Achsenabschnitt und a die Steigung der Geraden darstellen. Die Steigung a gibt an, wie stark der Anstieg von y pro Einheit von x ist. Wenn a positiv ist, bedeutet dies, dass y zunimmt, wenn x zunimmt, und umgekehrt. Wenn a negativ ist, bedeutet dies, dass y abnimmt, wenn x zunimmt, und umgekehrt. Der y-Achsenabschnitt a gibt den Wert von y an, wenn x gleich Null ist.[1]

2.2 Least Squares Methode

Die Least Squares Methode (LS-Methode) ist eine Methode der linearen Algebra, die in der linearen Regression eingesetzt wird, um eine Schätzung der Koeffizienten in der Regressionsgleichung zu erhalten. Diese Methode kann für jede beliebige Funktionstyp angewendet. Die Least Squares Methode minimiert die Summe der Abweichungen zwischen den beobachteten y-Werten und den vorhergesagten y-Werten. Die vorhergesagten y-Werte werden aus der Regressionsgleichung berechnet. Die Abweichung zwischen den beobachteten Datenpunkten und den vorhergesagten Werten kann sowohl positive als auch negative Werte haben. Wenn man die Abweichungen einfach addieren würde, würden sich positive und negative Abweichungen gegenseitig aufheben und die Summe könnte nahezu null sein, auch wenn die tatsächliche Abweichung insgesamt groß ist. Um dieses Problem zu vermeiden und die Abweichungen angemessen zu erfassen, quadriert man die Abweichungen, bevor man sie addiert. Durch das Quadrieren der Abweichungen werden alle Werte positiv, und größere Abweichungen haben einen größeren Einfluss auf die Gesamtsumme der quadrierten Abweichungen. Das bedeutet, dass die LS-Methode stärker auf größere Abweichungen reagiert und versucht, diese zu minimieren. Mathematisch lässt sich die Least Squares Methode wie folgt formulieren: Angenommen, wir haben n Beobachtungen von Paaren $(x1, y1), (x2, y2), \ldots, (xn, yn)$. Hierfür sucht man eine

[1] (Rice, 2007)

Funktion $f(x)$, die den Zusammenhang zwischen x und y am besten beschreibt. Diese Funktion $f(x)$ besitzt n-beliebigen Koeffizienten, die man herausfinden möchte. $f(a_0, \ldots, a_n)$ ist somit die Quadratsumme:

$$f(a_0, \ldots, a_n) = \sum_{n=1}^{M} (y_n - f(x_n))^2$$

Formel 1-1

Hierzu wird partielle Ableitung genommen und gleich Null gestellt, um die Quadratsumme zu minimieren:

$$\frac{\partial}{\partial a_n} f(a_0, \ldots, a_n) = 0$$

Formel 1-2

Die LS-Methode ist somit zur Approximation verschiedener Funktionstypen fähig, wie zum Beispiel eine Lineare Funktion.[2]

2.3 Lineare Least Squares Methode

Die Linear Least Squares Methode (LLS-Methode) ist eine spezielle Anwendung der Least Squares Methode, die in der linearen Regression verwendet wird, um die Regressionskoeffizienten zu schätzen. Die Linear Least Squares Methode funktioniert nur für lineare Modelle, das heißt Modelle, bei denen der Zusammenhang zwischen den Variablen durch eine lineare Funktion beschrieben werden kann. Angenommen, wir haben n Beobachtungen von Paaren $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, und wir suchen eine lineare Regressionsgleichung der Form $\hat{y} = a + bx$, die den Zusammenhang zwischen x und y

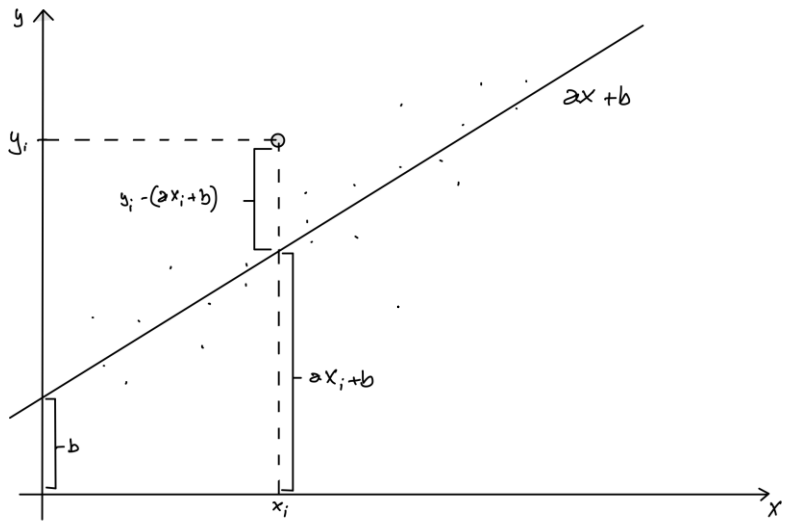

Abb. 1

[2] (Hermann, 2011)

am besten beschreibt (Abb. 1). Das $\hat{y}$ steht für das approximierter y-Wert. Man verwendet die LLS-Methode, um die Werte von a und b zu finden, in dem man die Summe der quadrierten Abweichungen zwischen den beobachteten y-Werten y_n und den vorhergesagten y-Werten $a + bx_n$ minimiert. Um die Gesamtsumme der quadratischen Abweichungen für alle Datenpunkte zu berechnen, addieren wir die quadrierten Abweichungen über alle M Datenpunkte:

$$f(a,b) := \sum_{n=1}^{M} \big(y_n - (ax_n + b)\big)^2$$

Formel 1-3

Durch Minimierung der Abweichungsfunktion $f(a,b)$ kann man die besten Schätzungen für die Parameter a und b erhalten, die die lineare Funktion am besten an die gegebenen Datenpunkte anpassen. Dies wird durch partielles Ableiten der Abweichungsfunktion nach a und b und das Gleichsetzen der Ableitungen mit Null erreicht, um die optimalen Werte für a und b zu finden.

$$\frac{\partial f}{\partial a} = 0, \qquad \frac{\partial f}{\partial b} = 0$$

Formel 1-4

Leitet man nach dem Koeffizienten a ab bekommt man folgende Gleichung:

$$\frac{\partial f}{\partial a} = \sum_{n=1}^{M} 2\big(y_n - (ax_n + b)\big) \cdot (-x_n) = 0 \qquad \text{I}$$

Formel 1-5

Leitet man nach dem Koeffizienten b ab bekommt man folgende Gleichung:

$$\frac{\partial f}{\partial b} = \sum_{n=1}^{M} 2(y_n - (ax_n + b)) \cdot (-1) = 0 \qquad \text{II}$$

Formel 1-6

Aus diesem beiden Gleichungen bekommt man ein Gleichungssystem, das wie folgt ausgedrückt werden kann:

$$\begin{pmatrix} \sum_{n=1}^{M} {x_i}^2 & \sum_{n=1}^{M} x_i \\ \sum_{n=1}^{M} x_i & (M+1) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{M} x_n y_n \\ \sum_{n=1}^{M} y_n \end{pmatrix}$$

Formel 1-7

Löst man das Gleichungssystem, ergibt sich folgende Formel für den Koeffizienten a:

$$a = \frac{(M+1)\left(\sum_{n=1}^{M} x_n y_n\right) - \left(\sum_{n=1}^{M} x_n\right)\left(\sum_{n=1}^{M} y_n\right)}{(M+1)\left(\sum_{n=1}^{M} {x_n}^2\right) - \left(\sum_{n=1}^{M} x_n\right)^2}$$

Formel 1-8

Und für den Koeffizient b:

$$b = \frac{\left(\sum_{n=1}^{M} {x_n}^2\right)\left(\sum_{n=1}^{M} y_n\right) - \left(\sum_{n=1}^{M} x_n y_n\right)\left(\sum_{n=1}^{M} x_n\right)}{(M+1)\left(\sum_{n=1}^{M} {x_n}^2\right) - \left(\sum_{n=1}^{M} x_n\right)^2}$$

Formel 1-9

Insgesamt zeigt die Herleitung der LLS-Methode, dass es eine Gerade approximiert, die am besten zu den Daten passt. Durch die Minimierung der quadratischen Fehler zwischen den tatsächlichen und approximierten Werten ermöglicht sie eine präzise Annäherung an Funktionen. Die LLS-Methode bietet somit einen soliden mathematischen Rahmen für die effektive Analyse und Modellierung von Daten, was sie zu einer wertvollen Technik in verschiedenen Bereichen der Wissenschaft und Technik macht.[3]

3 Fehleranalyse, Validierung und Alternativen der LLS-Methode

Die Fehleranalyse und Validierung der Linearen Least Squares Methode sind wichtige Schritte, um die Zuverlässigkeit und Genauigkeit der Ergebnisse dieser Methode zu überprüfen. Dazu braucht man verschiedene Werkzeuge bzw. Formeln, die für die Analyse und Validierung notwendig sind. Zum einem gibt es die Stichprobenvarianz der unabhängigen Variablen x. Diese Formel (Formel 2-1) subtrahiert den Durchschnittswert ($\bar{x}$) aller x-Werte von jedem einzelnen x-Wert, quadriert das Ergebnis und summiert alle quadrierten Abweichungen. Schließlich wird das Ergebnis durch die Anzahl der Beobachtungen geteilt, um die durchschnittliche quadratische Abweichung zu erhalten.

$$S_{xx} = \frac{1}{M}\sum_{i=0}^{M}(x_i - \bar{x})^2$$

Formel 2-1

Zum anderem kann man auch die Stichprobenvarianz der abhängigen Variablen y berechnen. Diese Formel (Formel 2-2) ähnelt der Formel für S_{xx}, nur dass sie auf die y-Werte angewendet wird. Der Durchschnittswert der y-Werte ($\bar{y}$) wird von jedem einzelnen y-Wert subtrahiert, das Ergebnis wird quadriert und alle quadrierten Abweichungen werden

[3] (Miller, 2012)

summiert. Schließlich wird das Ergebnis durch die Anzahl der Beobachtungen geteilt, um die durchschnittliche quadratische Abweichung zu erhalten.

$$S_{yy} = \frac{1}{M}\sum_{i=0}^{M}(y_i - \bar{y})^2$$

Formel 2-2

Außerdem gibt noch die Formel (Formel 2-3), die die Kovarianz zwischen den Variablen x und y berechnet. Sie misst, wie sich die x- und y-Werte gemeinsam von ihren jeweiligen Durchschnittswerten ($\bar{x}$ und $\bar{y}$) unterscheiden. Die Formel multipliziert die Abweichungen der x-Werte von ihrem Durchschnitt mit den Abweichungen der y-Werte von ihrem Durchschnitt und summiert diese Produkte über alle Beobachtungen. Das Ergebnis wird durch die Anzahl der Beobachtungen geteilt, um den durchschnittlichen Wert zu erhalten.

$$S_{xy} = \frac{1}{M}\sum_{i=0}^{M}(y_i - \bar{y})\,(x_i - \bar{x})$$

Formel 2-3

Diese Formeln sind alleine aussageschwach und sind insbesondere für die Berechnung von R^2 und r in diesem Kapitel notwendig.

3.1 Residuenanalyse

Die Residuenanalyse ist ein wichtiger Schritt bei der linearen kleinsten Quadrate (LLS)-Methode zur Parameterabschätzung in der linearen Regression. Sie ermöglicht die Überprüfung der Modellanpassung an die Daten und die Identifizierung von systematischen Abweichungen in den Residuen. Die Residuen sind die Differenzen zwischen den beobachteten abhängigen Variablenwerten und den vorhergesagten Werten des linearen Modells:

$$\epsilon_i = y_i - \hat{y}_i$$

Formel 2-4

Man kann aus der Beobachtung der Punktwolke und der Berechnung der Residuen zur Regressionslinie eine Aussage treffen, ob es um einen linearen Sachverhalt handelt.

3.1.1 Ausreißer

Ausreißer sind Datenpunkte, die deutlich von den anderen Datenpunkten in einem Datensatz abweichen. Sie sind ungewöhnlich oder im Vergleich zu den übrigen Beobachtungen stark abweichen und können potenziell Einfluss auf die Ergebnisse einer statistischen Analyse haben. Ausreißer können verschiedene Gründe haben, wie z.B. Messfehler,

Datenverunreinigungen, unerwartete Ereignisse oder tatsächliche Abweichungen in den Daten. Sie können sowohl zufällig als auch systematisch auftreten. Man kann Ausreißer wie folgt behandeln. Man kann sie einfach entfernen, wenn sie als fehlerhaft oder nicht repräsentativ angesehen werden. Außerdem können Sie durch Logarithmus- oder Wurzeltransformation reduziert oder abgeschwächt werden. Wenn es sich vermutlich um Messfehler oder Datenverunreinigungen handelt, könnte man die Daten (wenn möglich) nochmal ermitteln.[4]

3.2 Korrelationskoeffizient r

Ein weiterer Wert für Einschätzung des linearen Zusammenhangs ist der Korrelationskoeffizient. Er wird berechnet, indem die Kovarianz durch die Quadratwurzel des Produkts der Stichprobenvarianzen S_{xx} und S_{yy} dividiert wird (Formel 2-5). Der Korrelationskoeffizient gibt an, wie stark die lineare Beziehung zwischen den Variablen x und y ist. Ein Wert von -1 bedeutet eine perfekte negative Korrelation (negative Steigung), ein Wert von 1 eine perfekte positive Korrelation (positive Steigung) und ein Wert von 0 deutet auf keine lineare Korrelation hin.[5]

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Formel 2-5

3.3 Bestimmtheitsmaß R²

Das Bestimmtheitsmaß, auch als Bestimmtheitsgrad oder R²-Wert bezeichnet, wird verwendet, um die Güte der Anpassung eines Regressionsmodells an die Daten zu bewerten. Er gibt an, wie gut die Variation der abhängigen Variable durch das Modell erklärt wird.

R² ist das Verhältnis zwischen der Varianz zwischen $\hat{y}$ und $\bar{y}$ und der S_{yy}. Daraus ergibt sich folgende Formel:

$$R^2 = \frac{\frac{1}{M}\sum_{i=0}^{M}(\hat{y}_i - \bar{y})^2}{\frac{1}{M}\sum_{i=0}^{M}(y_i - \bar{y})^2}$$

Formel 2-6

Ein R²-Wert von 1 bedeutet, dass das Modell die gesamte Variation der abhängigen Variable erklärt, während ein R²-Wert von 0 darauf hinweist, dass das Modell keine Variation erklärt.

[4] (Baltes-Götz, 2022)
[5] (D.C.Agnew/C.Constable, 2008)

3.4 Andere Methoden

Obwohl die LLS-Methode weit verbreitet ist, gibt es alternative Methoden, die möglicherweise in bestimmten Situationen besser geeignet sind. Ein solcher Ansatz ist die robuste Regression, die Ausreißer besser berücksichtigt und eine stabilere Koeffizienten Schätzung ermöglicht. Die LLS-Methode hingegen basiert auf der Annahme, dass die Residuen normalverteilt sind, und ist daher empfindlich gegenüber Ausreißern oder Verletzungen dieser Annahme.[6] Ein weiterer alternativer Ansatz ist die Bayesianische Regression, bei der die Unsicherheit in den Schätzungen der Regressionskoeffizienten berücksichtigt wird. Dies wird erreicht, indem prior-Wahrscheinlichkeiten für die Koeffizienten festgelegt und dann mit den beobachteten Daten aktualisiert werden. Durch die probabilistische Modellierung kann die Unsicherheit in den Schätzungen besser erfasst werden.[7] Des Weiteren gibt es die Möglichkeit auf andere Abhängigkeit, statt die Lineare, zu überprüfen, die besser zu den Daten passt und besser erklären kann. Dies kann mit der LS-Methode für beliebige Funktionstypen oder Polynome angegangen werden.

4 Implementierung von Linear Least Squares

Die Implementierung der Linear Least Squares Methode in einer Programmiersprache ermöglicht die praktische Anwendung dieser Methode auf reale Daten. Python, eine beliebte Programmiersprache in der Datenanalyse und maschinellen Lernverarbeitung, bietet eine breite Palette von Bibliotheken und Funktionen, die die Umsetzung der Linearen Least Squares Methode erleichtern.

4.1 Programm zur Berechnung von Linear Least Squares

Der im Anhang beigefügte Code verwendet die Python-Bibliothek Matplotlib, um ein Streudiagramm und eine lineare Regression zu erstellen. Zunächst werden die benötigten Bibliotheken importiert: „matplotlib.pyplot“ für die Diagrammerstellung, „numpy“ für numerische Berechnungen, sowie „csv“ zum Lesen der Daten aus der CSV-Datei. Das Diagrammfenster wird mit einer Größe von 6x6 Zoll erstellt, und ein Koordinatensystem hinzugefügt. Die Achsen des Diagramms werden mit den Bezeichnungen "x-axis" und "y-axis" versehen. Es werden Variablen A, B, C und D initialisiert, die für die Berechnung der Koeffizienten der linearen Regression verwendet werden. Diese Variablen werden zunächst mit 0 initialisiert.

[6] (Jann, 2010)
[7] (Bishop, 2006)

$$A = \sum_{i=1}^{M} x_i y_i \; ; \; B = \sum_{i=1}^{M} x_i \; ; \; C = \sum_{i=1}^{M} y_i \; ; \; D = \sum_{i=1}^{M} {x_i}^2$$

Formel 3-1

Daraus folgt die folgende Formel für die Berechnung der Koeffizienten a und b (wie bei Formel 1-8 und Formel 1-9):

$$a = \frac{(M+1) \cdot A - B \cdot C}{(M+1) \cdot D - B^2}$$

Formel 3-2

$$b = \frac{D \cdot C - B \cdot A}{(M+1) \cdot D - B^2}$$

Formel 3-3

Danach wird der Benutzer aufgefordert, den Dateinamen der CSV-Datei einzugeben. Die Daten werden aus der CSV-Datei gelesen und in den Listen x und y gespeichert. Dabei wird das Semikolon als Trennzeichen angegeben, da die Werte in der CSV-Datei durch Semikolons getrennt sind. Für jede Reihe werden die verschiedene Abschnitte berechnet, die für die lineare Regression benötigt werden. Während der Eingabe der Punkte werden auch die Variablen A, B, C und D aktualisiert, um die erforderlichen Summen für die Berechnung der linearen Regression zu speichern. Die Punkte werden im Koordinatensystem mit ax.scatter(x, y) dargestellt. Anschließend werden die Koeffizienten a und b für die lineare Regression berechnet. Es wird eine Linie der linearen Regression generiert, indem 100 x-Werte im Bereich der minimalen und maximalen x-Werte erzeugt werden. Mit den berechneten Koeffizienten a und b wird die entsprechende y-Koordinate für jeden x-Wert berechnet $\hat{y} = a * x + b$. Die Linie der linearen Regression wird mit ax.plot$(x, y, '-r')$ gezeichnet. Schließlich werden das Diagramm an die Daten angepasst, indem die Achsen automatisch skaliert werden (ax.autoscale()) und ein zusätzlicher Rand von 0,1 hinzugefügt wird (ax.margins(0.1)). Das fertige Diagramm wird angezeigt.

4.1.1 Bibliotheken

Für die Methode der Linearen Kleinsten Quadrate (LLS) gibt es verschiedene Bibliotheken in Python, die bei der Implementierung dieser Methode helfen können. Eine wichtige Bibliothek ist NumPy, die eine leistungsstarke Unterstützung für numerische Berechnungen in Python bietet. Man kann die Funktion „linalg.lstsq“ verwenden, die direkt die Koeffizienten a und b berechnet.[8] Für maschinelles Lernen bietet scikit-learn eine nützliche

[8] (num23)

Unterstützung. Obwohl sie hauptsächlich für maschinelles Lernen entwickelt wurde, enthält die Bibliothek Funktionen für lineare Regression und LLS. Die Klasse sklearn.linear_model.LinearRegression ermöglicht die Anwendung der LLS-Methode auf Datenpunkte.[9] Diese genannten Bibliotheken bieten umfassende Funktionalitäten für die LLS-Methode in Python und erleichtern die Implementierung und Berechnung von linearen Anpassungen. Auch wenn Excel keine Bibliothek ist, kann man es trotzdem zur Berechnung und Visualisierung der Ausgleichsgerade verwenden.

5 Anwendungen der Linearen Least Squares Methode

Dieses Kapitel beschäftigt sich mit der Anwendung der linearen Least-Squares-Methode auf reale Daten zur Untersuchung der Abhängigkeit der Reaktionsgeschwindigkeit von der Konzentration von Thiosulfat. Die Reaktionsgeschwindigkeit ist ein wichtiger Parameter in der Chemie und kann uns wertvolle Informationen über die Geschwindigkeit und den Verlauf chemischer Reaktionen liefern. Die Konzentration von Thiosulfat, einem weit verbreiteten chemischen Stoff, wird als unabhängige Variable betrachtet, während die Reaktionsgeschwindigkeit als abhängige Variable gemessen wird.

5.1 Datenbeschaffung und -vor Verarbeitung

Für die Untersuchung der Abhängigkeit der Reaktionsgeschwindigkeit von der Konzentration (C) von Thiosulfat wurden experimentelle Daten gesammelt. Die Datenbeschaffung umfasste die Durchführung einer Reihe chemische Reaktionen unter verschiedenen Konzentrationen von Thiosulfat.

Nr.	Vol. S_2O_3-Lsg. In ml	Vol. H_2O in ml	Vol. H+ in ml	C	v[1/t]s^-1	t in s
	0	0	0	0	0	0
1	50	0	5	0,1364	0,03466	28,85
2	40	10	5	0,1091	0,02857	35,43
3	30	20	5	0,0818	0,0185	53,93
4	20	30	5	0,0545	0,0114	88,06
5	10	40	5	0,0273	0,0059	169,74

Tabelle 1

Zu Beginn wurden die benötigten Chemikalien und Geräte für die Experimente vorbereitet. Eine Lösung von Thiosulfat wurde hergestellt, wobei verschiedene Konzentrationen sorgfältig gemessen und vorbereitet wurden. Zusätzlich wurden andere Reagenzien und Lösungen vorbereitet, die für die Durchführung der Reaktion erforderlich waren.

[9] (sci231)

Die Experimente wurden unter kontrollierten Bedingungen durchgeführt, wobei die Reaktionszeit für jede Konzentration von Thiosulfat beim Farbumschlag gemessen wurde. Die Reaktionsgeschwindigkeit wurde dementsprechend berechnet. Daraus folgt Tabelle 1.

5.2 Einfache Lineare Regression und Analyse

Um die lineare Regression für die vorliegenden Daten durchzuführen und die Beziehung zwischen der Konzentration (x-Achse) und der Reaktionsgeschwindigkeit (y-Achse) zu analysieren, verwenden man die Methode der kleinsten Quadrate (LLS). Das Ziel ist es, eine lineare Funktion zu finden, die die Daten am besten approximiert. Hierfür wird den im Anhang beigefügten Code verwendet. Bei der Ausgabe in der Konsole ergibt sich eine Steigung von etwa a ≈ 0,26 und ein y-Achsenabschnitt von b ≈ -0,0008. Die Steigung a bedeutet, dass man einen Anstieg von ungefähr 0,26 Einheiten in der Reaktionsgeschwindigkeit für jede Einheit Zunahme in der Konzentration erwarten können. Der y-Achsenabschnitt b von -0,0008 gibt den geschätzten Wert der Reaktionsgeschwindigkeit an, wenn die Konzentration 0 ist. Somit ergibt sich die folgende Gleichung für die Regressionslinie:

$$y = 0{,}26x - 0{,}0008$$

Diese Gleichung ermöglicht es, die Reaktionsgeschwindigkeit in Abhängigkeit von der Konzentration vorherzusagen. Um zu beurteilen, wie gut das Modell zu den Daten passt, berechnen man den Grad der Bestimmtheit (R^2). In diesem Fall ist R^2 = 99,05 %. Zusätzlich wurde ein Korrelationskoeffizient von r = 0,9952 ermittelt, um die Stärke und Richtung der linearen Beziehung zwischen Konzentration und Reaktionsgeschwindigkeit zu beurteilen.

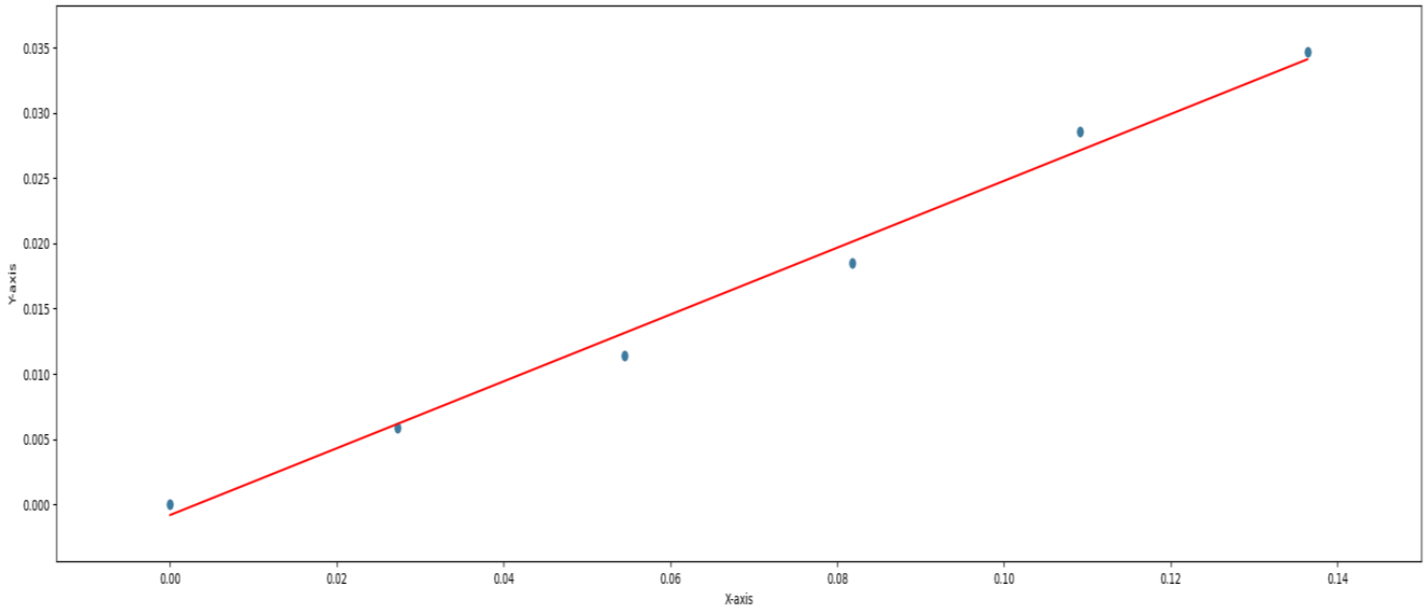

Abb. 2

5.3 Interpretation der Ergebnisse

Bei Betrachtung der Punktwolke (Abb. 2) erkennt man, dass es keine Ausreißer gibt und dass die Regressionslinie nahezu perfekt zu den Punkten passt. Die Residuen betragen ungefähr ± 0.0001, was darauf hindeutet, dass der Chemieversuch gut abgelaufen ist und die Daten echt sein sollten. Ein Korrelationskoeffizient von 0,9952 weist auf eine sehr starke positive lineare Beziehung zwischen Konzentration und Reaktionsgeschwindigkeit hin. Dies weist darauf hin, dass eine Konzentrationserhöhung mit einem deutlichen Anstieg der Reaktionsgeschwindigkeit einhergeht. Werte nahe 1 zeigen an, dass die Datenpunkte sehr nahe an der Regressionsgeraden liegen, dass die Beziehung linear und sehr gut ist. Das Bestimmtheitsmaß (R^2) von 99,05 % zeigt an, dass das lineare Modell gut zu den Daten passt und dass die Konzentration einen sehr starken Einfluss auf die Reaktionsgeschwindigkeit hat. Eine hohe Bestimmtheit (R^2) deutet darauf hin, dass die Gleichung die beobachtete Variation gut erklärt und Vorhersagen der Reaktionsgeschwindigkeit basierend auf der Konzentration mit hoher Genauigkeit getroffen werden können. Zusammenfassend lässt sich sagen, dass die Analyse der Punktwolke und die statistischen Kennzahlen darauf hindeuten, dass der Chemieversuch erfolgreich verlaufen ist und die Konzentration linear zur Reaktionsgeschwindigkeit abhängig ist. Allerdings gibt es ein Problem, und zwar dass eine Konzentration von 0 mol pro ml keine Reaktion hervorruft und somit auch keine negative Reaktionsgeschwindigkeit möglich ist. In diesem Fall kann die Methode der kleinsten Quadrate (LS-Methode) auf die Formel $f(x) = ax$ angewendet werden, da der y-Achsenabschnitt bei null liegt, und um einen Wert für a zu erhalten. In diesen Fall ist die Steigung $a = 0{,}2478$. Dabei wird der R^2-Wert und der r-Wert etwas niedriger ausfallen.

6 Fazit

Im Rahmen dieser Facharbeit wurde intensiv die Anwendung der "Linear Least Squares"-Methode für die einfache lineare Regression untersucht. Die einfache lineare Regression ist eine grundlegende statistische Technik zur Untersuchung des linearen Zusammenhangs zwischen einer abhängigen Variable (y) und einer unabhängigen Variable (x). Die Methode der "Linear Least Squares" erwies sich dabei als äußerst effektiv und gängige Methode, um den besten linearen Zusammenhang zwischen den Variablen zu approximieren. Ein wesentlicher Vorteil dieser Methode liegt in ihrer Einfachheit und intuitiven Interpretierbarkeit, was quantitative Aussagen über den linearen Zusammenhang zwischen den Variablen ermöglicht. Die Interpretation der Koeffizienten ist von großer Bedeutung, da sie Einblicke in Richtung und Stärke des linearen Zusammenhangs geben.

Es ist jedoch wichtig zu beachten, dass die Validität der Ergebnisse von bestimmten Annahmen und Fehleranalysen abhängt. Eine wesentliche Annahme der Methode der einfachen linearen Regression ist die Linearität des Zusammenhangs zwischen den Variablen. Es wird angenommen, dass die Beziehung zwischen der abhängigen und der unabhängigen Variable durch eine Gerade repräsentiert werden kann. Diese Annahme kann durch den Bestimmtheitsgrad und den Korrelationskoeffizienten überprüft werden. Zudem sollten Ausreißer genauer analysiert werden, um die Robustheit der Ergebnisse sicherzustellen. Falls diese Annahmen verletzt sind, müssen alternative Regressionsmethoden wie die nicht lineare oder Polynom-Regression in Betracht gezogen werden. Zusätzlich sollten weitere Tests durchgeführt werden, um die Gültigkeit der Ergebnisse zu überprüfen, beispielsweise die Prüfung der Punktwolke, um Muster oder Anomalien in den Daten zu identifizieren. In Bezug auf die Anwendung der LLS-Methode auf reale Daten zur Untersuchung der Abhängigkeit der Reaktionsgeschwindigkeit von der Konzentration von Thiosulfat konnte gezeigt werden, dass die Methode gut geeignet ist, um den linearen Zusammenhang zu analysieren und Vorhersagen zu treffen. Die Regressionsanalyse ergab eine Regressionsgerade mit einer Steigung von etwa 0,26 und einem y-Achsenabschnitt von etwa -0,0008. Die Werte des Bestimmtheitsgrads (R^2 = 99,05%) und des Korrelationskoeffizienten (r = 0,9952) zeigten eine sehr starke positive lineare Beziehung zwischen Konzentration und Reaktionsgeschwindigkeit. Die Analyse der Punktwolke und statistischen Kennzahlen unterstützten diese Ergebnisse.

Für die zukünftige Entwicklung dieser Methode und als Forschungsziel könnten weitere Tests und Analysen berücksichtigt werden, um die Gültigkeit der Ergebnisse weiter zu überprüfen. Es wäre interessant zu untersuchen, wie die LLS-Methode im R^3 oder mit mehrere unabhängige Variablen angewendet werden kann. Außerdem könnte die Methode mit anderen statistischen Techniken oder Modellen kombiniert werden, um die Vorhersagegenauigkeit zu verbessern.

Abschließend lässt sich sagen, dass die Methode der einfachen linearen Regression mit der LLS-Methode ein äußerst wertvolles Werkzeug ist, um den linearen Zusammenhang zwischen Variablen zu analysieren und Vorhersagen zu treffen. Es besteht ein Bedarf an einer weiteren Entwicklung und Diskussion dieser Methode im schulischen Kontext, um den Schülerinnen und Schülern ein fundiertes Verständnis für diese statistische Methode zu vermitteln. Es ist jedoch wichtig, die Annahmen angemessen zu verstehen und ihre Gültigkeit sorgfältig zu überprüfen, um zuverlässige Schlussfolgerungen zu ziehen.

7 Anhang

7.1 Literaturverzeichnis

- *numpy.org.* [Online] [Zitat vom: 16. Mai 2023.] https://numpy.org/doc/stable/reference/generated/numpy.linalg.lstsq.html#numpy.linalg.lstsq.
- *scipy.org.* [Online] [Zitat vom: 16. Mai 2023.] https://docs.scipy.org/doc/scipy/reference/generated/scipy.linalg.lstsq.html.
- *scikit-learn.org.* [Online] [Zitat vom: 16. Mai 2023.] https://scikit-learn.org/stable/modules/linear_model.html.
- **Baltes-Götz, Bernhard. 2022.** *Lineare Regressionsanalyse mit SPSS.* uni-trier : s.n., 2022.
- **Bishop, Christopher M. 2006.** *Learning, Pattern Recognition And Machine.* New York : Springer, 2006. ISBN 0-387-31073-8.
- **D.C.Agnew/C.Constable. 2008.** igpphome.ucsd.edu. [Online] 2008. https://igpphome.ucsd.edu/~cathy/Classes/SIO223A/sio223a.chap7.pdf.
- **Hermann, Martin. 2011.** *Numerische Mathematik. 3.* München: Oldenbourg : s.n., 2011. ISBN: 978-3-486-70820-2.
- **Jann, Ben. 2010.** Springer. [Online] 2010. https://link.springer.com/chapter/10.1007/978-3-531-92038-2_27.
- **Miller, Steven J. 2012.** The Method of Least Squares. *Department of Mathematics and Statistics.* Brown University : s.n., 2012.
- **Rice, John A. 2007.** Linear Least Squares. *Mathematical Statistics and Data Analysis.* University of California, Berkeley : Thomson Brooks/Cole, 2007.
- **studyflix.** studyflix. [Online] [Zitat vom: 16. Mai 2023.] https://studyflix.de/statistik/bestimmtheitsmas-2146.

7.2 Code

```
import matplotlib.pyplot as plt
import numpy as np
import csv
# Fenstergröße und Erstellung
fig = plt.figure(figsize=(6, 6))
# Erstellung Koordinatensystem
ax = fig.add_subplot(111, aspect='equal')
# x und y label
ax.set_xlabel('X-axis')
ax.set_ylabel('Y-axis')
A = float(0)
B = float(0)
C = float(0)
D = float(0)
# Angabe der CSV-Datei
csv_file = input('Gib den Dateinamen der CSV-Datei ein: ')
# Daten aus CSV-Datei lesen
x = []
y = []
with open(csv_file, 'r') as file:
    reader = csv.reader(file, delimiter=';')  # Trennzeichen angeben
    next(reader)  # Überspringe die Header-Zeile
    for row in reader:
        x_val = float(row[0])
        y_val = float(row[1])
        x.append(x_val)
        y.append(y_val)
        # Berechnung der verschiedenen Abschnitte
        # A; B; C; D
        A = float(x_val * y_val + A)
        B = float(x_val + B)
        C = float(y_val + C)
        D = float(x_val * x_val + D)
# Punkte im Koordinatensystem anzeigen
ax.scatter(x, y)
# a / b -koeffizient
M = len(x)
a = float(((M + 1) * A - B * C) / ((M + 1) * D - (B * B)))
b = float((D * C - A * B) / ((M + 1) * D - (B * B)))
print(a)
print(b)
x = np.linspace(min(x), max(x), 100)
y = a * x + b
ax.plot(x, y, '-r')
# Diagramm an Daten anpassen
ax.autoscale()
ax.margins(0.1)

# Diagramm anzeigen
plt.show()
```

Erklärung über die selbständige Anfertigung der Arbeit

Ich erkläre, dass ich die vorliegende Facharbeit ohne fremde Hilfe angefertigt und nur die im Literaturverzeichnis angeführten Quellen und Hilfsmittel verwendet habe.

Bad Ems, 12.06.2023

_____

Adnan Aidak