Lösung zu Aufgabe 10: Lineare Regression

Bei der Anwendung der linearen Regression wird durch eine Menge von Wertepaaren (x_i, y_i) nach der Methode der geringsten quadratischen Abweichungen eine Gerade gelegt. Die allgemeine Ansatzfunktion einer Geraden lautet:

$$y = bx + a$$

Darin ist b der sogenannte Regressionskoeffizient, welcher die Steigung der Geraden angibt und a ist der Achsenabschnitt. Da eine nach der Methode der kleinsten Fehlerquadrate ermittelte Ausgleichsgerade stets durch den Schwerpunkt $(\overline{x}, \overline{y})$ der Einzelpunkte verläuft, kann bei Kenntnis des Regressionskoeffizienten b der Achsenabschnitt durch Einsetzen des Schwerpunktes berechnet werden:

$$a = \overline{y} - b\overline{x}$$

Ersetzen wir in der allgemeinen Ansatzfunktion der Geraden den Achsenabschnitt a entsprechend, so erhalten wir die gängige Darstellung der Regressionsgeraden, bei welcher der Achsenabschnitt nicht explizit sondern implizit durch den Schwerpunkt $(\overline{x}, \overline{y})$ dargestellt wird:

$$(y - \overline{y}) = b(x - \overline{x})$$

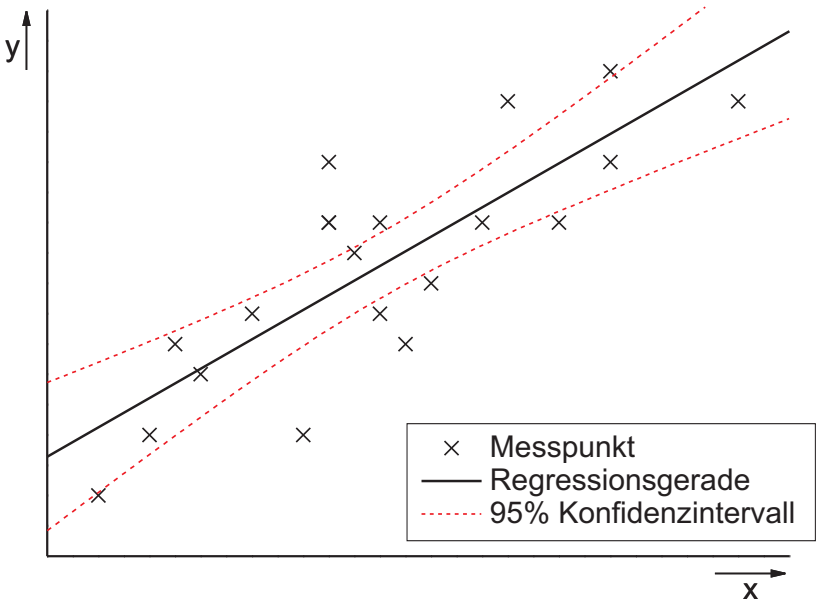
Der beste Schätzwert für den Regressionskoeffizient b wird wie folgt berechnet:

$$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Es ist sinnvoll, diese Berechnung mit Hilfe des in den meisten wissenschaftlichen Taschenrechnern verfügbaren Statistikmodus vorzunehmen. Möchte man die Berechnung hingegen ohne diese Unterstützung vornehmen, reduziert die folgende, mathematisch äquivalente Darstellung der Gleichung den Rechenaufwand erheblich:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

Dass es sich bei der so errechneten Geraden nur um einen Schätzwert handelt, verdeutlicht nachfolgende Abbildung. Da sowohl die Berechnung des Steigungsmaßes b als auch der Schwerpunkt $(\overline{x}, \overline{y})$ basierend auf abweichungsbehafteten Messdaten erfolgt, sind über die Abweichungsfortpflanzung sowohl die Position als auch die Steigung der Geraden mit Abweichungen behaftet. Das resultierende Konfidenzintervall weist, wie in der Abbildung eingetragen, im Bereich des Schwerpunktes $(\overline{x}, \overline{y})$ seine geringste Breite auf und erweitert sich zu beiden Seiten mit zunehmendem Abstand.



Es stellt sich daher, ähnlich wie bei anderen bereits behandelten statistischen Verfahren, die Frage nach der Unsicherheit dieser Schätzung. Ziel der vorliegenden Aufgabe ist daher zunächst, zu dem nach obiger Gleichung berechneten Regressionskoeffizienten b ein zugehöriges Konfidenzintervall zu berechnen. Das Konfidenzintervall für den Regressionskoeffizienten b zur statistischen Sicherheit $P = 1 - \alpha$ beträgt:

$$\left[b - \frac{\hat{\sigma} t_{n-2;1-\alpha/2}}{\sqrt{n} S_x}, \quad b + \frac{\hat{\sigma} t_{n-2;1-\alpha/2}}{\sqrt{n} S_x} \right]$$

Neben dem p-Quantil der Student'schen t-Verteilung $t_{n-2;1-\alpha/2}$, der empirischen Streuung der x-Werte S_x sowie dem Stichprobenumfang n hängt das Konfidenzintervall demnach von der sogenannten Restvarianz $\hat{\sigma}^2$ ab, welche wie folgt berechnet wird:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{j=1}^{n} \left(y_j - \overline{y} + b(\overline{x} - x_j) \right)^2$$

Bei dem darin enthaltenen Term $\left(y_j - \overline{y} + b(\overline{x} - x_j) \right)$ handelt es sich anschaulich jeweils um die Differenz zwischen einem beobachteten y-Wert y_j und dem Funktionswert der Regressionsgeraden an der Stelle x_j.

a) Bestimmung des Empfindlichkeitskoeffizienten a_c mittels linearer Regression

In der vorliegenden Aufgabe soll die lineare Regression genutzt werden, um auf der Basis von Messdaten den thermischen Empfindlichkeitskoeffizienten a_c eines Kondensators zu ermitteln. Hierzu wird folgender funktionaler Zusammenhang angegeben, welcher die tatsächliche Kapazität C mit dem Empfindlichkeitskoeffizienten a_c, der Nennkapazität C_0 sowie der Temperatur T und einer Referenztemperatur T_0 in Bezug setzt:

$$C = C_0 \cdot \left(1 + a_c \cdot \left(T - T_0 \right) \right)$$

Um die weitere Berechnung möglichst einfach zu gestalten, ist es sinnvoll, obige Gleichung derart umzustellen, dass wir eine Struktur erhalten, bei welcher der Koeffizient a_c formal identisch mit dem Regressionskoeffizienten b ist. Im vorliegenden Fall bietet sich folgende Umformung an:

$$\frac{C}{C_0} - 1 = a_c \cdot (T - T_0)$$

Vergleichen wir diese Struktur mit der allgemeinen Geradengleichung erkennen wir folgende Zusammenhänge:

$$\frac{C}{C_0} - 1 = y$$

$$T - T_0 = x$$

$$a_c = b$$

Prinzipiell sind auch davon abweichende Umformungen der gegebenen Gleichung möglich und sinnvoll. Es ist jedoch darauf zu achten, dass die neu konstruierte x-Größe von der unabhängigen Größe im Rahmen der Versuchsdurchführung abhängt. Aus der Versuchsbeschreibung ergibt sich im vorliegenden Fall, dass es sich bei der Temperatur T um die unabhängige x-Größe handelt, da diese jeweils gezielt eingestellt wird. Die zugehörige Kapazität des Kondensators stellt in diesem Fall folglich die von x abhängige y-Größe dar.

Nach obigen Zusammenhängen können wir nun aus den in der Aufgabenstellung gegebenen Werten unsere x-y-Wertepaare für die nachfolgende Regressionsberechnung ermitteln. Mit $C_0 = 68\,\text{nF}$, $T_0 = 20°\text{C}$ und den Messwerten für T und C aus der Tabelle erhalten wir folgende x-y-Wertepaare:

i	1	2	3	4	5	6	7	8
x / °C	-20	-15	-10	-5	0	5	10	15
y / 1	0,049118	0,036912	0,025294	0,012353	0,000735	-0,011912	-0,024412	-0,037941

Um aus diesen Werten den Regressionskoeffizienten berechnen zu können, bestimmen wir zunächst die Mittelwerte $\overline{x}$ und $\overline{y}$:

$$\overline{x} = -2{,}5°\text{C}$$

$$\overline{y} \approx 6{,}268 \cdot 10^{-3}$$

Mit diesen Werten ergibt sich der Regressionskoeffizient zu:

$$b \approx -2{,}474 \cdot 10^{-3}\,\frac{1}{°\text{C}}$$

Die Unsicherheit c_b dieser Schätzung, also die Breite des Konfidenzintervalls ergibt sich gemäß obiger Gleichung mit:

$$c_b = \frac{\hat{\sigma} t_{n-2;1-\alpha/2}}{\sqrt{n} S_x}$$

Um das Konfidenzintervall des Regressionskoeffizienten bestimmen zu können, berechnen wir zunächst die Restvarianz $\hat{\sigma}^2$:

$$\hat{\sigma}^2 \approx 3{,}283 \cdot 10^{-7}$$

Damit ergibt sich die für das Konfidenzintervall benötigte Standardabweichung $\hat{\sigma}$ zu:

$$\hat{\sigma} \approx 5{,}73 \cdot 10^{-4}$$

Das darüber hinaus benötigte p-Quantil der Student'schen t-Verteilung $t_{n-2;1-\alpha/2}$ ergibt sich mit n = 8 und α = 0,05 zu:

$$t_{n-2;1-\alpha/2} = t_{6;0{,}975} = 2{,}447$$

Weiterhin benötigen wir noch die Streuung S_x der x-Werte. Diese errechnet sich zu:

$$S_x \approx 12{,}2475°C$$

Damit können wir die Unsicherheit c_b berechnen:

$$c_b = \frac{5{,}73 \cdot 10^{-4} \cdot 2{,}447}{\sqrt{8} \cdot 12{,}2475°C} \approx 4{,}05 \cdot 10^{-5} \frac{1}{°C}$$

Das vollständige Messergebnis für den Empfindlichkeitskoeffizienten a_c, welcher identisch ist mit dem hier berechneten Regressionskoeffizienten b, ergibt sich damit zu:

$$a_c = b = -2{,}474 \cdot 10^{-3} \frac{1}{°C} \pm 4{,}05 \cdot 10^{-5} \frac{1}{°C} \;;\; P = 95\%$$

b) Resultierende Kapazität bei einer bestimmten Temperatur T einschließlich Vertrauensbereich

Die in Aufgabenteil a) mittels linearer Regression berechnete Gerade ordnet einem beliebig gewählten x-Wert x^* einen entsprechenden y-Wert y^* zu, für den gilt:

$$y^* = \bar{y} + b(x^* - \bar{x})$$

Aufgrund der Unsicherheit der Geraden selbst ist auch die auf dieser basierende Zuordnung von x- und y-Werten mit einer Unsicherheit behaftet. Ein Konfidenzintervall für y^* zur statistischen Sicherheit P = 1 – α beträgt:

$$\left[y^* - \frac{\hat{\sigma} t_{n-2;1-\alpha/2}}{\sqrt{n}} \sqrt{1 + \frac{(x^* - \bar{x})^2}{S_x^2}},\quad y^* + \frac{\hat{\sigma} t_{n-2;1-\alpha/2}}{\sqrt{n}} \sqrt{1 + \frac{(x^* - \bar{x})^2}{S_x^2}} \right]$$

In der Aufgabenstellung wird der spezielle x-Wert x^* nicht direkt angegeben, sondern indirekt über die Temperatur $T^* = 27{,}5°C$. Für unsere Regressionsberechnung hatten wir x-Werte zugrunde gelegt, die mit der Temperatur T wie folgt zusammen hängen:

$$x = T - T_0$$

Für den gesuchten Wert x^* ergibt sich daher:

$$x^* = T^* - T_0 = 27{,}5°C - 20°C = 7{,}5°C$$

Der zugehörige y-Wert y^* errechnet sich damit zu:

$$y^* = \bar{y} + b(x^* - \bar{x}) = 6{,}268 \cdot 10^{-3} - 2{,}474 \cdot 10^{-3} \frac{1}{°C} \cdot (7{,}5°C + 2{,}5°C) \approx -0{,}01847$$

Für die Berechnung des zugehörigen Konfidenzintervalls benötigen wir noch das p-Quantil der Student'schen t-Verteilung zu den Parametern n = 8 und $\alpha = 0{,}02$. Hierfür gilt:

$$t_{n-2;1-\alpha/2} = t_{6;0{,}99} = 3{,}143$$

Die Unsicherheit c_{y^*} ergibt sich mit den vorliegenden Zahlenwerten zu:

$$c_{y^*} = \frac{5{,}73 \cdot 10^{-4} \cdot 3{,}143}{\sqrt{8}} \sqrt{1 + \frac{(7{,}5°C + 2{,}5°C)^2}{(12{,}2475°C)^2}} \approx 8{,}22 \cdot 10^{-4}$$

Das vollständige Messergebnis für y^* lautet damit:

$$y^* = -0{,}01847 \pm 8{,}22 \cdot 10^{-4} \;;\; P = 98\%$$

Da in der Aufgabenstellung nach einem vollständigen Messergebnis für die Kondensatorkapazität C gefragt ist, handelt es sich hierbei jedoch noch nicht um das Endergebnis. Wie wir in Aufgabenteil a) festgestellt haben, gilt für y und C folgender Zusammenhang:

$$y = \frac{C}{C_0} - 1$$

Der Mittelwert der von y^* abhängigen Kapazität C^* kann zunächst einfach durch Umstellen obiger Gleichung und Einsetzen der Zahlenwerte bestimmt werden. Für den in Aufgabenteil b) betrachteten Kondensator mit einem Nennwert von $C_0 = 47$ nF gilt:

$$C^* = C_0 \cdot (y^* + 1) = 47\ \text{nF} \cdot (-0{,}01847 + 1) \approx 46{,}132\ \text{nF}$$

Da es sich bei der Nennkapazität C_0 um eine Konstante handelt, hängt die gesuchte Unsicherheit der Kapazität nur von der Unsicherheit von y ab. Formal können wir diesen Zusammenhang durch eine Abweichungsfortpflanzungsrechnung mit nur einer abweichungsbehafteten Eingangsgröße ermitteln. Für die zugehörige Unsicherheit c_{C^*} gilt gemäß den bekannten Zusammenhängen für die Fortpflanzung zufälliger Abweichungen allgemein:

$$c_{C^*} = \sqrt{\left(\left.\frac{\partial C}{\partial y}\right|_{y^*} \cdot c_{y^*} \right)^2}$$

Die partielle Ableitung von C nach y an der Stelle y^* ergibt sich zu:

$$\left.\frac{\partial C}{\partial y}\right|_{y^*} = C_0 = 47\ \text{nF}$$

Damit ergibt sich die Unsicherheit c_{C^*} zu:

$$c_{C^*} = \sqrt{\left(47\ \text{nF} \cdot 8{,}22 \cdot 10^{-4}\right)^2} \approx 0{,}0386\ \text{nF}$$

Das vollständige Messergebnis der Kondensatorkapazität C bei einer Temperatur von T = 27,5°C beträgt somit:

$$C^* = 46{,}132\ \text{nF} \pm 0{,}0386\ \text{nF}\ ;\ P = 98\%$$