(1983)

FLÄCHEN UNTER REGRESSIONSKURVEN

nach R. Suich und H. Rutemiller, California State University, Fullerton
Originaltitel in 'Teaching Statistics' Vol. 4 (1982) Nr. 1: Areas unter Regression Curves
Übersetzung: I. Strauß, Bearbeitung: B. Wollring

Zu den nützlichsten und interessantesten Gegenständen, die man in einem Statistik-Kurs vorstellen kann, gehört die Behandlung der auf Quadratsummen-Minimierung basierenden Regressionsgeraden. Sie sind nicht nur von großem historischem Interesse, sondern haben auch viele Anwendungen auf allen möglichen Gebieten.

Im vorliegenden Beitrag zeigen wir eine einfache Erweiterung. Sie wird wichtig, wenn man nicht den Graphen selbst, wohl aber die Fläche unter diesem betrachten will. Die Diskussion über diese Erweiterung sollte die Schüler und Studenten dazu bringen, gründlicher darüber nachzudenken, wozu die Regressionsgeraden tauglich sind und wozu nicht.

Wir geben hier davon aus, daß die Entwicklung der Regressionsgeraden der Lerngruppe bereits bekannt ist. Folgende Bezeichnungen werden wir benutzen:

x: unabhängige Variable
y: abhängige Variable
$\hat{y}(x) = mx + b$: Gleichung der Regressionsgeraden, erstellt durch Minimieren der Quadratsumme, mit Steigung m und y-Achsenabschnitt b

Die dargestellte Erweiterung kann irgendwann nach der Einführung erfolgen. Das Interesse der Autoren wurde zum ersten Mal während der Beschäftigung mit Kontroll-Messungen zur Luft-Verschmutzung geweckt, und dieser 'Aufhänger' ist auch in die folgende Darstellung eingearbeitet.

Das Problem

Wir wollen die Kohlenmonoxid-Emission zweier PKW P_1 und P_2 testen. Dazu messen wir die Verunreinigungs-Rate (CO in Gramm pro Meile) bei beiden Wagen. Angenommen, für P_1 ergibt sich 50 und 40 für P_2. Aufgeschreckt vom schlechten Abschneiden ihres Produktes, erklärt die Herstellerfirma von P_1, daß eine einmalige Messung kein getreues Abbild der Verhältnisse über einen längeren Zeitraum darstellt. Sie besteht deshalb auf einer Meßreihe in 1000 Meilen-Intervallen für die ersten 10 000 Meilen. Solche Registrierungen werden tatsächlich etwa alle 1000 Meilen für jedes Auto durchgeführt und sind in Tabelle 1 aufgelistet.

PKW P_1

x	0	1	2	3	4.2	5	6	6.9	8	9.2	10
Y	50	56	58	60	58	63	73	71	76	73	80

PKW P_2

x	0	1.1	2.2	3	4	5.3	6	7	8.1	9	10
y	40	49	58	65	75	77	86	93	98	103	109

Tabelle 1 : CO-Emission für die PKW P_1 und P_2. x bezeichnet die Fahrstrecke in Einheiten von 1000 Meilen, y die CO-Emission in Gramm pro Meile

Zur Auswertung der Tabelle dienen folgende Daten:

$$\text{PKW } P_1: \quad \sum x_i = 55.3 \quad ; \sum y_i = 718.0 \quad ; \sum x_i y_i = 3918.0$$
$$\sum x_i^2 = 388.89 \quad ; \sum y_i^2 = 47808.0$$
$$\bar{x} = 5.027 \quad ; \sum (x_i - \bar{x})^2 = 110.88$$

$$\text{PKW } P_2: \quad \sum x_i = 55.7 \quad ; \sum y_i = 853.0 \quad ; \sum x_i y_i = 5062.4$$
$$\sum x_i^2 = 390.75 \quad ; \sum y_i^2 = 71283.0$$
$$\bar{x} = 5.064 \quad ; \sum (x_i - \bar{x})^2 = 108.71$$

Ein Blick auf Tabelle 1 zeigt, daß P_1 bei den ersten Messungen höhere Emissionswerte aufweist, während er später vergleichsweise besser abschneidet. Wir stehen also vor dem Problem, wie die beiden Datenmengen verglichen werden können. Eine Methode besteht darin, die beiden Regressionsgeraden $\hat{y}(x) = mx + b$ zu berechnen. Wir erhalten:

PKW P_1 :

$$m = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n} = 2.78$$

$$b = \bar{y} - m\bar{x} = 51.28$$

$$\hat{y}(x) = 2.78x + 51.28$$

PKW P_2 : $\hat{y}(x) = 6.84x + 42.93$

Die Regressionsgeraden sind in Bild 1 dargestellt.

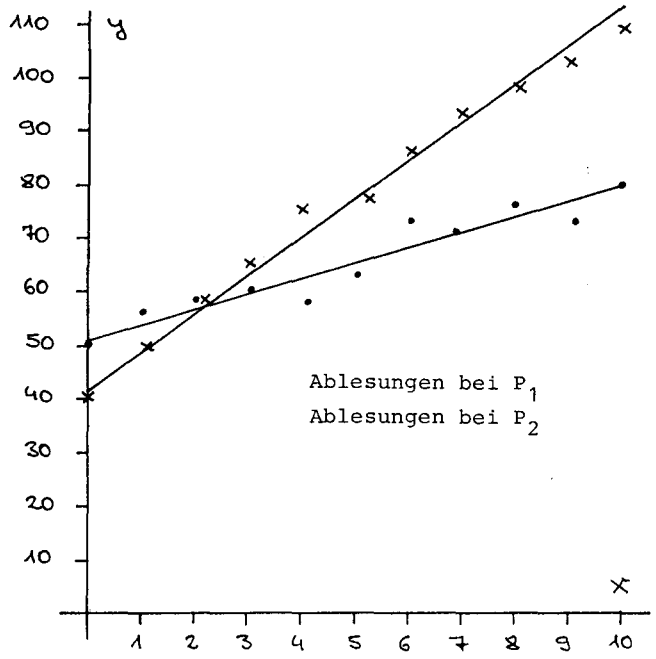

Bild 1 : Regressionsgeraden bei P_1 und P_2

Tabelle 2 zeigt die Ergebnisse der Varianzanalyse zu beiden Meßreihen (vergleiche z.B. KREYSZIG): Dabei ist $Q_2/(n-2)$ die Varianz der Verteilung aller Abweichungen der y_i von den Funktionswerten $\hat{y}(x_i)$.

	Varianzanalyse bei P_1			Varianzanalyse bei P_2		
	Freiheitsgrade	Quadratsumme	Durchschnittsquadrat	Freiheitsgrade	Quadratsumme	Durchschnittsquadrat
Regression	1	858.422	858.422	1	5080.007	5080.007
Residuum	9	83.759	9.307	9	56.719	6.302
Total	10	942.181		10	5136.727	

Legende

	Freiheitsgrade	Quadratsumme	Durchschnittsquadrat
Regression	1	$Q_1 = Q - Q_2$	Q_1
Residuum	n - 2	$Q_2 = \sum (y_i - \hat{y}(x_i))^2$	$Q_2/(n-2)$
Total	n - 1	$Q = \sum (y_i - \bar{y})^2$	

Tabelle 2 : Varianzanalyse zu den Meßreihen bei P_1 und P_2

Es mag wünschenswert sein, diese Ausrechnungen als Übungsaufgaben zu stellen und die Lernenden zu fragen, wie sie die Emissionen der beiden Autos vergleichen würden. Man beachte, daß die einfache Gerade eine gute Anpassung an die Meßwerte darstellt.

Nun stellt sich die Frage: "Wie können wir die Emissionswerte über die gesamte Betrachtungszeit hin vergleichen?"

Die Antwort werden vielleicht einige Ihrer Schüler selbst finden: Wir sind ja nicht an einem Vergleich der Emissionsraten zu einem bestimmten Zeitpunkt interessiert, d.h. wir wollen nicht die $\hat{y}$-Werte für ein bestimmtes x vergleichen. Vielmehr suchen wir den Vergleich der gesamten Emission bei beiden Wagen über die gesamte 10 000-Meilen-Strecke, wollen also die Flächen unter den Regres-

sionsgeraden vergleichen. Die Trapezfläche unter der Geraden zu P_1 beträgt $0.5 \cdot 10\,000 \cdot (51.3 + 79.1) = 652$, die bei der Geraden zu P_2 ist $0.5 \cdot 10\,000 \cdot (42.9 + 111.3) = 771$. Die gesamte Emission von P_1 während der 10 000 Meilen ist tatsächlich geringer als die von P_2.

Die Diskussion in der Lerngruppe kann an dieser Stelle problemlos durch einige wenige Bemerkungen abgeschlossen werden. Diese sollten den Hinweis einschließen, daß obige Daten nur Stichproben aus einem 10 000 Meilen-Intervall darstellen und somit nicht notwendigerweise eine wirkliche Differenz bei den beiden Wagen signalisieren. Will man die Diskussion noch fortsetzen, bietet sich - in Ergänzung zur Signifikanzfrage - die Erörterung des Problems an, ob verschiedene Exemplare der Typen P_1 und P_2 getestet werden sollten. Und zusätzlich: Erhält man dieselben Ergebnisse in aufeinanderfolgenden Drehzahlbereichen? Fuhr ein oder fuhren zwei Fahrer die beiden Wagen? In welchen Drehzahlbereichen liegt P_2 günstiger als P_1?

Die zugehörige Test-Theorie

Wir wollen den Zusammenhang zwischen x und y mit dem linearen Ansatz $y_i = \tilde{m}x + \tilde{b} + \varepsilon_i$ beschreiben. Dabei sehen wir die ε_i als beobachtete Werte von unabhängigen normalverteilten Zufallsgrößen mit Erwartungswert 0 und Varianz σ^2 an. Wir interessieren uns für den Wert A der Fläche unter der Regressionsgeraden von x = 0 bis x = 10 (·1000) :

$$A = \int_0^{10} (\tilde{m}x + \tilde{b})\,dx = 50\tilde{m} + 10\tilde{b}$$

Wir benutzen für $\tilde{m}$ und $\tilde{b}$ die üblichen Schätzfunktionen, die die Quadratsumme minimieren, und erhalten als beobachtete Fläche unter der Regressionsgeraden a = 50m + 10b. Dabei sind m und b beobachtete Werte der Zufallsgrößen:

$$M = \frac{\sum (x_i - \bar{x})\,Y_i}{\sum (x_j - \bar{x})^2} \qquad \text{und} \qquad B = \bar{Y} - \bar{x}M$$

Und a ist beobachteter Wert der Zufallsgröße A = 50M + 10B. Dabei sind die Y_i unabhängige normalverteilte Zufallsgrößen mit Erwartungswert $\tilde{m}x_i + \tilde{b}$ und Varianz σ^2. Für A gilt nun:

$$\begin{aligned} A &= 50M + 10B \\ &= (50 - 10\bar{x})M + 10\bar{Y} \\ &= \sum_i \left[\frac{(50 - 10\bar{x})(x_i - \bar{x})}{\sum (x_j - \bar{x})^2} + \frac{10}{n} \right] Y_i \\ &= \sum \alpha_i Y_i \end{aligned}$$

Als Linearkombination der unabhängigen normalverteilten Zufallsgrößen Y_i ist auch A normalverteilt, und nach dem Additionssatz für Varianzen folgt (nach länglicher Umformung):

$$\begin{aligned} \text{Var}\,(A) &= \sum \alpha_i^2 \sigma^2 \\ &= \frac{50^2 n - 2 \cdot 50 \cdot 10 n\bar{x} + 100 \sum x_i^2}{n \sum (x_j - \bar{x})^2} \cdot \sigma^2 \end{aligned}$$

Nach Einsetzen der entsprechenden Werte erhalten wir:

$$\begin{aligned} \text{Var}\,(A_1) &= \frac{2500 \cdot 11 - 2 \cdot 50 \cdot 10 \cdot (-55.3) + 100 \cdot 388.97}{11 \cdot 110.97} \cdot \sigma^2 \\ &= 9.091\,\sigma^2 \end{aligned}$$

Eine entsprechende Rechnung für A_2 ergibt $\text{Var}\,(A_2) = 9.094\,\sigma^2$. Da A_1 und A_2 voneinander unabhängig sind, folgt:

$$\begin{aligned} \text{Var}\,(A_1 - A_2) &= \text{Var}(A_1) + \text{Var}\,(A_2) = 9.091\,\sigma^2 + 9.094\,\sigma^2 \\ &= 18.185\,\sigma^2 \end{aligned}$$

Ferner ist $A_1 - A_2$ als Linearkombination unabhängiger normalverteilter Zufallsgrößen wieder normalverteilt mit dem Erwartungswert $E(A_1 - A_2) = E(A_1) - E(A_2)$ und der oben genannten Varianz.

Wir wollen die Hypothese testen, daß P_1 und P_2 in dem 10 000 - Meilen-Intervall gleiche Emissionen haben. Das bedeutet $H_0 : E(A_1 - A_2) = 0$ gegen $H_1 : E_1(A_1 - A_2) \neq 0$. Da σ^2 nicht bekannt ist, schätzen wir es durch folgende Größe:

$$\frac{1}{2}\left(\begin{array}{c}\text{Durchschnittsquadrat} \\ \text{des Residuums bei } P_1\end{array} + \begin{array}{c}\text{Durchschnittsquadrat} \\ \text{des Residuums bei } P_2\end{array}\right)$$

$$= \frac{9.307 + 6.302}{2}$$

$$= 7.804$$

Wir wenden den t-Test an: Ist $E(A_1-A_2) = 0$, so ist die Prüfgröße

$$T = \frac{(A_1 - A_2) - 0}{\sqrt{\text{Schätzwert von } Var(A_1-A_2)}}$$

t-verteilt mit 18 Freiheitsgraden (siehe Tabelle 2). Wir erhalten als beobachteten Wert:

$$t = \frac{652 - 771}{\sqrt{18.185 \cdot 7.804}} = - 9.99$$

Dieses Resultat ist hochsignifikant: Eine Tabelle zur t-Verteilung zeigt, daß bei 18 Freiheitsgraden $P(t \leq -3.61) \leq 0.001$ gilt. Wir schließen daher, daß zwischen den Emissionsmengen von P_1 und P_2 in dem beobachteten 10 000 - Meilen-Intervall ein hochsignifikanter Unterschied besteht.

Verallgemeinerung

Das beschriebene Verfahren kann mit Hilfe der Matrizen-Technik für jede Polynomfunktion in x verwendet werden.

Literatur

DRAPER, N.; SMITH, H.: Applied Regression Analysis. - New York: John Wiley and sons; second edition

GRAYBILL, F.A.: Theory and Application of the Linear **Model**. - North Scituate, Mass.: Duxbury Press

KREYSZIG, E.: Statistische Methoden und ihre Anwendungen. - **Göttingen**: Vandenhoeck und Ruprecht 1979[7]

STORM, R.: Wahrscheinlichkeitsrechnung, mathematische **Statistik** und Qualitätskontrolle. - Leipzig: VEB Fachbuchverlag 1976[6]

YAMANE, T.: Statistik. Ein einführendes Lehrbuch. - **Fischer** Taschenbuch 1976