EIN HÄUFIGER PATZER BEI EINFACHER LINEARER REGRESSIONSANALYSE

von Keith Sandrock, University of the Witwatersrand, Johannesburg, South Africa
Originaltitel in "Teaching Statistics" Vol.10 (1988) Nr.3: A Common Pitfall of Simple linear Regression Analysis
Übertragung: Bernd Wollring, Universität Münster

Lineare Regressionsanalyse ist ein sehr beliebtes analytisches Werkzeug zur Untersuchung von Zusammenhängen. Aus diesem Grund enthält sogar fast jeder billige Taschenrechner fest verdrahtete Funktionen zur linearen Regression. Man hat damit ein äußerst nützliches Hilfsmittel zum Testen von Hypothesen oder für die Interferenzstatistik. Es ist ein Teil im Rahmen des Gesamtkonzeptes "Lineare Modelle".

Das Kriterium, auf das man sich im allgemeinen am meisten stützt, um zu entscheiden, ob ein Modell angemessen ist oder nicht, ist der Wert des Korrelationskoeffizienten r.

Der Leser sei zu folgendem kleinen Experiment eingeladen: Man frage jemanden, der sich laufend mit linearer Regressionsanalyse befaßt, welches Resultat er als erstes anschaut, den Wert des Achsenabschnittes a, oder den Wert der Steigung b oder den Wert von r. Die Antwort wird fast ausschließlich sein: "Ich betrachte zunächst r, denn daran sehe ich, wie gut oder schlecht das Modell ist."

Es ist allerdings eine Fallgrube, sein Urteil mit dem Wert r zu begründen. Das folgende Beispiel soll dieses illustrieren.

Es ist aus zwei Gründen ein nützliches Beispiel. Zunächst gibt es sehr wenige Datenpaare. Daher kommt das übliche gute Verfahren, eine Skizze der Residuen nach DRAPER und SMITH (1981) zu untersuchen, nicht in Frage. Zum zweiten liegen

die Daten ganz offensichtlich linear und geben dem Analytiker keinen rechten Anlaß, die Hypothese

$$Y_i = a + bt_i + e_i \quad \text{where } e_i \sim N(0, \sigma^2) \tag{1}$$

zurückzuweisen.

Das Beispiel

Das Ergebnis eines bestimmten Prozesses ist abhängig von der Temperatur. Mit dem Ansteigen der Temperatur sinken die Produktionskosten in Rand/Tonne. Es ist sehr teuer, Pilotstudien durchzuführen, besonders bei höheren Temperaturen, daher stehen für die Analyse nur fünf Meßpunkte zur Verfügung. Sie sind in Tabelle 1a dargestellt. Die Frage lautet: "Bei welcher Temperatur muß das Verfahren durchgeführt werden, um die Kosten pro Tonne zu minimieren?"

Tabelle 1a : Die Daten des Beispiels

Temperatur (^{o}C)	10	20	30	40	50
Kosten pro Tonne	125.1	116.3	109.2	104.6	102.0

Tabelle 1b : Ergebnisse zum Beispiel

Ergebnisse der Regressionsanalyse

Achsenabschnitt	: a = 128.1	zur Schätzung von α
Steigung	: b = - 0.579	zur Schätzung von β
Korrelationskoeffizient	: r = - 0.978	

F-Test

Berechneter Wert : F = 63.3
Tabellenwert auf 5% Niveau = 10.13
Ergebnis: Eine signifikante Regression, die anzeigt, daß das lineare Modell die Quadratsumme der Fehler signifikant verringert.

Signifikanz von r

Berechneter t-Wert : -9.2
Tabellenwert auf 5% Niveau mit 3 Freiheitsgraden = 3.182

Ergebnis: Der Wert von r ist signifikant. Dies bedeutet eine sehr gute lineare Korrelation von Temperatur und Kosten.

95%-Konfidenzintervalle für α und β

$P\ (121.14 \leq \alpha \leq 136.48) = 0.95$
$P\ (-0.81 \leq \beta \leq -0.35) = 0.95$

Schlußfolgerung: Die Analyse gibt keinen Anlaß zum Zweifel an der Hypothese (1).

Eine lineare Regressionsanalyse erscheint in der Tat sehr zufriedenstellend. Der Wert $r = -0.977$ und die üblichen statistischen Ergebnisse, die eine solche Analyse umfaßt, geben dem Untersuchenden keinen Anlaß die lineare Hypothese (1) zurückzuweisen.

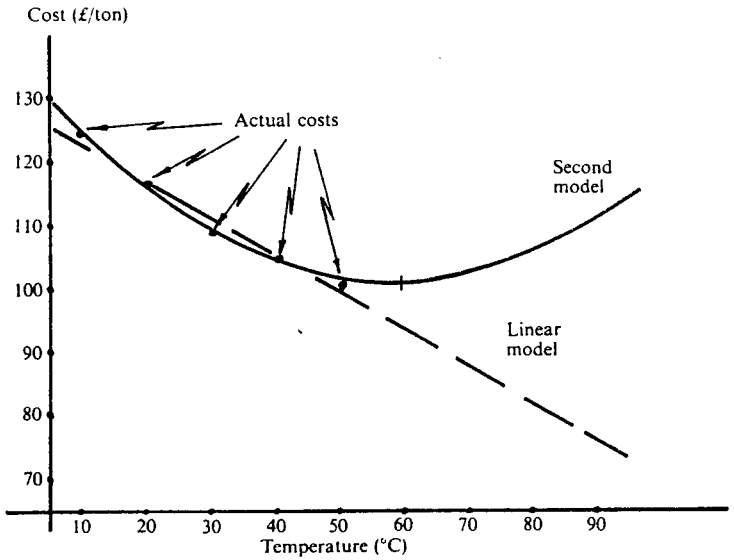
Damit ist die Antwort auf die oben gestellte Frage einfach, den Vorgang bei dem größtmöglichen praktisch erreichbaren Temperaturwert zu fahren, in diesem Fall um 100°C , denn dort tritt anscheinend das Minimum der Kosten auf.

An diesem Punkt ist es jedoch instruktiv, die Annahme der Linearität erneut zu überprüfen. Eine Betrachtung der Punktwolke in Bild 1 legt den Verdacht einer Krümmung nahe. Eine Zeichnung mit Temperaturen von 0°C bis 60°C und Kosten zwischen 100 Rand und 130 Rand zeigt diese Krümmung deutlicher.

Bild 1 Numerische Daten zum Beispiel

Daten									
Temperatur	10	20	30	40	50	60	70	80	90
Kosten									
Tatsächl.Kosten	125.1	116.3	109.2	104.6	102.0	unbekannte Kosten			
Lineares Modell	123.0	117.2	111.4	105.7	99.9	94.1	88.3	82.5	76.7
Zweites Modell	125.4	115.8	109.5	104.1	103.0	101.4	102.8	106.2	111.6

Bild 1 : Graphik zum Beispiel



Minimale Kosten 101Rand/Tonne bei Temperatur 61°C

Dieses Phänomen kann man ohne Analysis einfach dadurch prüfen, daß man von jedem beobachteten Wert den der vorhergehenden Beobachtung subtrahiert und das Ergebnis durch die zugehörige Temperaturdifferenz dividiert. Bei dieser kleinen Übung entsteht eine Serie von Quotienten folgender Form:

$$\Delta C_i/\Delta T_i \qquad (2)$$

Dabei bedeuten $\Delta C_i = C_i - C_{i-1}$ = Abnahme der Koster bei Anstieg der Temperatur von T_{i-1} auf T_i und $\Delta T_i = T_i - T_{i-1}$.

Die Quotienten (2) sind in der letzten Spalte von Tabelle 2 dargestellt. Das Ergebnis ist ganz deutlich, denn gäbe es eine lineare Beziehung zwischen den Daten, so wären die Quotienten in etwa konstant und würden zufällig um ihren Mittelwert schwanken, wie in (1) gefordert. Dies trifft

offensichtlich nicht zu. Tatsächlich scheint gerade das Gegenteil der Fall zu sein, denn es ist klar, daß die Quotienten einen bestimmten Trend haben. Steigt die Temperatur, so fällt die Wachstumsrate bei den Kosten, und wir können darauf spekulieren, daß der Effekt dieser Krümmung auf einen Vorzeichenwechsel hinführt. Mit anderen Worten, die Kosten beginnen wieder zu steigen, wenn die Temperatur einen gewissen Wert überschreitet.

Tabelle 2

Temperatur in °C	Kosten C_i in Rand/Tonne	ΔKosten/Δ T
10	125.1	
20	116.3	- 0.88
30	109.2	- 0.71
40	104.6	- 0.46
50	102.0	- 0.26

Der nächste Schritt besteht darin, die Bestimmung der Differenzenquotienten mit den Werten der zweiten Spalte in Tabelle 2 anstelle der Kosten erneut durchzuführen. Man erhält drei Quotienten "zweiter Ordnung", wie Tabelle 3 zeigt. Nun scheinen die Werte einigermaßen konstant zu sein, ihr Mittelwert ist 0.207.

Tabelle 3

Temperatur in °C	Kosten C_i in Rand/Tonne	ΔKosten/ΔT	Quotienten zweiter Ordnung
10	125.1		
20	116.3	-0.88	
30	109.2	-0.71	0.17
40	104.6	-0.46	0.25
50	102.0	-0.26	0.20

Hat man nun stationäre Quotienten erreicht, so kann man ein alternatives Modell entwickeln. Die neue Hypothese ist:

$$(C_i - C_{i-1})/\Delta T_i - (C_{i-1} - C_{i-2})/\Delta T_{i-1} = \text{constant} + e_i \sim N(0, \sigma^2)$$

Also folgt:

$$\hat{C}_i = (\text{constant})\,\Delta T_i + (\Delta T_i/\Delta T_{i-1})(C_{i-1} - C_{i-2}) + C_{i-1}$$

Wählt man ferner die Werte der unabhängigen Variablen T äquidistant so gilt $\Delta T_i = \Delta T_{i-1} = \ldots = \Delta T$, und es folgt:

$$\hat{C}_i = \text{constant} + 2C_{i-1} - C_{i-2} \qquad (3)$$

Im hier vorgelegten Beispiel gilt:

$$\hat{C}_i = 2{\cdot}07 + 2C_{i-1} - C_{i-2} \qquad (4)$$

Gleichung (3) beschreibt einen autoregressiven Prozess (BOX und JENKINS, 1970) in seiner einfachsten Form. Allgemeiner notiert man ihn in der Form:

$$\hat{C}_i = (\text{constant}) + 2\varphi_0 C_{i-1} - \varphi_1 C_{i-2}$$

und er wird als AR(2)-Prozeß bezeichnet, da es sich um ein Modell zweiter Ordnung handelt.

Die aus Gleichung (4) erhaltenen Ergebnisse zeigt Bild 1. Sie unterscheiden sich wesentlich - alarmierend! - von denen des linearen Modells. Hier ist eine Entscheidung notwendig. An dieser Stelle muß man folgende Fragen beantworten:

- Welches der Modelle sollte man akzeptieren? Keines?
- Bei welcher Temperatur werden minimale Kosten erreicht?
- Wie groß sind die minimalen Kosten je Tonne?

Vor der Beantwortung dieser Fragen ist es notwendig, sich darüber klar zu werden, wie das Ergebnis (4) zustande kommt.

Zunächst wurden keine a-priori- Annahmen zur Linearität der Daten gemacht. Sie wurden nur zur Erzeugung von Quotienten erster und höherer Ordnung verwendet, bis diese stationär wurden. Die trat bei den Differenzquotienten zweiter Ordnung ein (Spalte 4,Tabelle 3). Man sollte hier mit statistischen Tests prüfen, ob stationäre Verhältnisse vorliegen, etwa mit einem χ^2-Test, aber das haben wir hier nicht aufgenommen. Man brauchte für einen realistischen χ^2-Test auch mehr Daten.

offensichtlich nicht zu. Tatsächlich scheint gerade das Gegenteil der Fall zu sein, denn es ist klar, daß die Quotienten einen bestimmten Trend haben. Steigt die Temperatur, so fällt die Wachstumsrate bei den Kosten, und wir können darauf spekulieren, daß der Effekt dieser Krümmung auf einen Vorzeichenwechsel hinführt. Mit anderen Worten, die Kosten beginnen wieder zu steigen, wenn die Temperatur einen gewissen Wert überschreitet.

Tabelle 2

Temperatur in °C	Kosten C_i in Rand/Tonne	ΔKosten/Δ T
10	125.1	
20	116.3	- 0.88
30	109.2	- 0.71
40	104.6	- 0.46
50	102.0	- 0.26

Der nächste Schritt besteht darin, die Bestimmung der Differenzenquotienten mit den Werten der zweiten Spalte in Tabelle 2 anstelle der Kosten erneut durchzuführen. Man erhält drei Quotienten "zweiter Ordnung", wie Tabelle 3 zeigt. Nun scheinen die Werte einigermaßen konstant zu sein, ihr Mittelwert ist 0.207.

Tabelle 3

Temperatur in °C	Kosten C_i in Rand/Tonne	ΔKosten/ΔT	Quotienten zweiter Ordnung
10	125.1		
20	116.3	-0.88	
30	109.2	-0.71	0.17
40	104.6	-0.46	0.25
50	102.0	-0.26	0.20

Hat man nun stationäre Quotienten erreicht, so kann man ein alternatives Modell entwickeln. Die neue Hypothese ist:

$$(C_i - C_{i-1})/\Delta T_i - (C_{i-1} - C_{i-2})/\Delta T_{i-1} = \text{constant} + e_i \sim N(0,\sigma^2)$$

Also folgt:

$$\hat{C}_i = (\text{constant})\,\Delta T_i + (\Delta T_i/\Delta T_{i-1})(C_{i-1} - C_{i-2}) + C_{i-1}$$

Wählt man ferner die Werte der unabhängigen Variablen T äquidistant so gilt $\Delta T_i = \Delta T_{i-1} = \ldots = \Delta T$, und es folgt:

$$\hat{C}_i = \text{constant} + 2C_{i-1} - C_{i-2} \qquad (3)$$

Im hier vorgelegten Beispiel gilt:

$$\hat{C}_i = 2{\cdot}07 + 2C_{i-1} - C_{i-2} \qquad (4)$$

Gleichung (3) beschreibt einen autoregressiven Prozess (BOX und JENKINS, 1970) in seiner einfachsten Form. Allgemeiner notiert man ihn in der Form:

$$\hat{C}_i = (\text{constant}) + 2\varphi_0 C_{i-1} - \varphi_1 C_{i-2}$$

und er wird als AR(2)-Prozeß bezeichnet, da es sich um ein Modell zweiter Ordnung handelt.

Die aus Gleichung (4) erhaltenen Ergebnisse zeigt Bild 1. Sie unterscheiden sich wesentlich - alarmierend! - von denen des linearen Modells. Hier ist eine Entscheidung notwendig. An dieser Stelle muß man folgende Fragen beantworten:

- Welches der Modelle sollte man akzeptieren? Keines?
- Bei welcher Temperatur werden minimale Kosten erreicht?
- Wie groß sind die minimalen Kosten je Tonne?

Vor der Beantwortung dieser Fragen ist es notwendig, sich darüber klar zu werden, wie das Ergebnis (4) zustande kommt.

Zunächst wurden keine a-priori- Annahmen zur Linearität der Daten gemacht. Sie wurden nur zur Erzeugung von Quotienten erster und höherer Ordnung verwendet, bis diese stationär wurden. Die trat bei den Differenzquotienten zweiter Ordnung ein (Spalte 4,Tabelle 3). Man sollte hier mit statistischen Tests prüfen, ob stationäre Verhältnisse vorliegen, etwa mit einem χ^2-Test, aber das haben wir hier nicht aufgenommen. Man brauchte für einen realistischen χ^2-Test auch mehr Daten.

Zum zweiten haben wir bei diesem Vorgehen Informationen ausgeschlossen, die die Zusammenhänge zwischen den beobachteten Werten der abhängigen Variablen betreffen: die Autokorrelationen. Autokorrelationen sind das Tor zu einer anderen Welt, einer Welt, die beschreibt, auf welche Art die Daten untereinander in Beziehung stehen. Als Analogie ist es hilfreich, Autokorrelationen und partielle Autokorrelationen in Bezug auf Daten als etwas Ähnliches anzusehen, wie das, was Gene für die lebende Welt sind. Sie legen die Charakteristik des Beobachteten fest, und dessen, was noch zukünftig beobachtet werden wird.

Zum dritten haben wir die unabhängige Variable, die Temperatur, in diesem Modell vernachlässigt. Dieses Modell behauptet nicht, daß die Kosten explizit eine Funktion der Temperatur sind, nur daß aufeinanderfolgende Kostenwerte untereinander in Beziehung stehen. Das Modell wurde durch die Datensätze selbst erzeugt. Es war nicht angenommen, daß zu den Daten ein lineares Modell gehört. Das korrekte Modell ist das Modell zweiter Ordnung (4). Minimale Kosten von 101 Rand enstehen bei der Temperatur 61°C.

Schlußfolgerung

An einer weit verbreiteten Technik zur Analyse von Daten hat sich ein Defizit verdeutlichen lassen, und der Leser ist auf eine Fallgrube aufmerksam gemacht worden. Hat man einen großen Datensatz zur Hand, so tut man gut daran, eine Skizze der Abweichungen herzustellen, um einen Anhaltspunkt dafür zu finden, daß die Hypothese (1) ggf. nicht zutrifft. Das Problem im täglichen Leben besteht oft darin, daß man unzureichende Daten hat, um informative Skizzen der Abweichungen zu erstellen. Ein anderes praktisches Problem besteht darin, daß der Untersuchende nicht imstande ist, die Werte der unabhängigen Variablen frei zu wählen. Er steht einem fait accompli gegenüber und muß so gut entscheiden, wie er kann. In dem vorgelegten Beispiel wäre ein einziger Punkt mit einer Temperatur über 50°C und den zugehörigen Kosten entscheidend gewesen.

Es hat sich herausgestellt, daß dennoch aufgrund der "mageren" Daten ein passendes Modell entwickelt werden konnte. Das ist die Gabe des Statistikers. Auf der Basis vollständiger Daten kann jeder Entscheidungen treffen. In dieser Hinsicht ist es nichts wert, daß es nicht notwendig zutrifft, daß eine große Datenmenge erforderlich ist, um Modelle höherer Ordnung zufriedenstellend anzupassen, vieles hängt an der Natur der Daten und der Art, wie sie gekoppelt sind.