KLÄRUNG DES KONZEPTS DER REGRESSION MIT HILFE VON DREI-PUNKT-DATENSÄTZEN

von LeRoy A. Franklin

Orginaltitel in "Teaching Statistics" Vol. 10 (1988) Nr. 1:

Clarifying Regression Concepts
Using 3 Point Data Sets

Übersetzung: Manfred Borovcnik, Klagenfurt

Kurzfassung: Zur Wiederholung und zur Vertiefung der Einsicht über die Grundlagen der Regression schlägt der Autor die Beschränkung auf nur 3 Punkte vor. Durch geschickte Wahl der Punkte läßt sich aufgrund der Variation der Daten die Bedeutung der wichtigsten Begriffe wie Varianzen, Residuen und Korrelationskoeffizient erklären.

1. Einleitung

Viele Anfängerstudenten finden die Berechnungen im Zusammenhang mit einfacher linearer Regression ganz simpel und klar. Leider werden viele ihrer charakteristischen Eigenheiten nicht wirklich verstanden und so mancher Student kommt erst nach mehreren Jahren einschlägiger Beschäftigung auf die Hintergründe solcher Konzepte wie dem "Bestimmungskoeffizienten" r^2. Der Autor hat herausgefunden, daß man die mühseligen Berechnungen minimieren kann, wenn man sich auf Drei-Punkt-Datensätze einschränkt. Wenn man jedoch diese drei Punkte sorgfältig auswählt, kann man sehr viel Einsicht in viele wichtige Regressionskonzepte gewinnen. Die folgenden sieben Beispiele sind vom Autor als *Rückblick* auf die Regressionsrechnung verwendet worden und setzen voraus, daß die Studenten bereits mit folgenden Konzepten und Definitionen konfrontiert worden sind (z.B. Ott, 1984):

Summe der Quadrate von x (Sum of Squares):

$$SS_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2$$

Summe der Quadrate von y:

$$SS_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{1}{n}(\Sigma y_i)^2$$

Summe der gemischten (quadratischen) Produkte:

$$SS_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \frac{1}{n}\Sigma x_i \Sigma y_i$$

Steigung der Regressionsgeraden:

$$\hat{b} = SS_{xy} / S_{xx}$$

Achsenabschnitt der Regressionsgeraden:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Summe der Quadrate, erklärt durch Regressionsbeziehung, das ist das Streuen der Ausgleichspunkte $\hat{y}_i = \hat{a} + \hat{b}x_i$ auf der Regressionsgeraden um den Mittelwert $\bar{y}$:

$$SS_{\hat{y}\hat{y}} = \Sigma(\hat{y}_i - \bar{y})^2 = \hat{b}SS_{xy}$$

Mittlerer quadratischer Fehler, das ist die"Restvarianz", die nicht erklärt wird durch Regression, das Streuen um die Regressionsgerade:

$$s^2 = SS_e/(n - 2) = \Sigma(y_i - \hat{y}_i)^2/(n - 2)$$

Es gilt folgende Streuungszerlegung:

$$SS_e = S_{yy} - S_{\hat{y}\hat{y}}$$

Bestimmungskoeffizient, das ist der Anteil der durch Regressionsbeziehung erklärten Varianz an der Gesamtvarianz der y-Daten bezogen auf $\bar{y}$, gleichzeitig das Quadrat des Korrelationskoeffizienten:

$$r^2 = S_{\hat{y}\hat{y}} / SS_{yy}$$

Test von H_o : b=O gegen H_1 : b>O, Prüfgröße

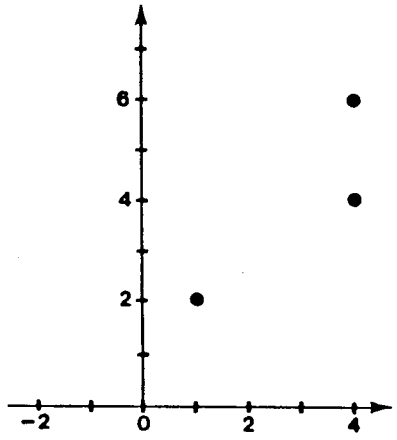
$$t_{n-2} = (\hat{b}-O) / \sqrt{s^2/SS_{xx}}$$

Geprüft wird, ob die Steigung der Regressionsgeraden signifikant von Null verschieden ist.

2. Sieben einfache Beispiele zur linearen Regression

Das erste Beispiel ist als Basisbeispiel zu verstehen, alle weiteren Beispiele werden damit verglichen werden.

Beispiel 1: Es geht um die Punkte (1,2), (4,4) und (4,6). Man beachte die Werte der verschiedenen Quadratsummen sowie den Umstand, daß die beste Ausgleichsgerade (im Sinne der Methode der kleinsten Quadrate) genau zwischen den Punkten (4,4) und (4,6) hindurchgeht, diese beiden Punkte also gleich gewichtet.

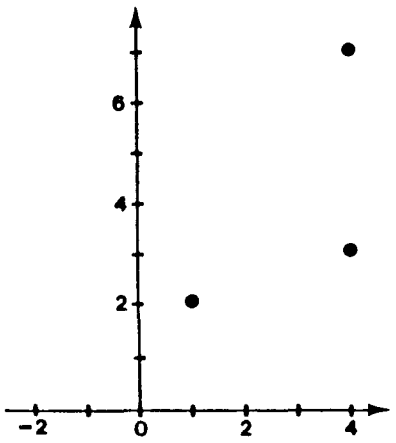


$SS_{xx} = 6$ $\quad \hat{\beta}_1 = 1$ $\quad R^2 = 0{\cdot}75$
$SS_{xy} = 6$ $\quad \hat{\beta}_0 = 1$ $\quad t = 1{\cdot}732$
$SS_{yy} = 8$ $\quad SS_{\hat{y}\hat{y}} = 6$ Signifikanz-
$\bar{x} = 3$ $\quad s^2 = 2$ niveau = 0·167
$\bar{y} = 4$

Beispiel 2: Die *beiden rechten Punkte* in Beispiel 1 *wandern gleichmäßig vertikal auseinander*: (1,2), (4,3), (4,7).
Es lohnt sich, die Studenten *vor* der Berechnung der Kennziffern zu fragen, welche Veränderung der Werte sie erwarten.
Die Regressionsgerade bleibt dieselbe wie in Beispiel 1, weil die Punkte symmetrisch auseinanderwandern, daher bleiben auch die Ausgleichspunkte $\hat{y}_i = \hat{a} + \hat{b}x_i$ gleich. Das hat zur Folge, daß die durch Regression erklärte Varianz (das Streuen der Ausgleichspunkte) $SS_{\hat{y}\hat{y}}$ gleich bleibt.
Weil die Punkte nun aber weiter von der Regressionsgeraden entfernt sind, wird die Restvarianz SS_e bzw. s^2 größer.
Die Punkte streuen mehr in der y-Koordinate, daher wird SS_{yy} größer.

Der Anteil der durch Regression erkärten Varianz, das ist $r^2 = SS_{\hat{y}\hat{y}} / S_{yy}$, wird demnach kleiner.
Die Prüfgröße für die Steigung b, das ist t_{n-2}, wird kleiner und weist nun eine geringere Signifikanz auf, obwohl der Wert von $\hat{b}$ sich *nicht* verändert hat.

Bemerkungen: Dasjenige Signifikanznivieau, auf dem die zu prüfende Nullhypothese für die vorliegenden Daten gerade noch abgelehnt werden kann, heißt *aktueller p-Wert* der Prüfgröße. Ist der aktuelle p-Wert 4,5 %, so heißt das, daß H_o für das Signifikanzniveau α = 5 % abgelehnt werden kann, für ein α = 4,4 % z.B. aber nicht mehr. Je *kleiner* der aktuelle p-Wert, desto *höher* signifikant der Wert der Prüfgröße. Im Fall unserer Prüfgröße t_{n-2} gilt: Je kleiner der Wert von t_{n-2}, desto größer der p-Wert, desto geringer die Signifikanz des Ergebnisses.

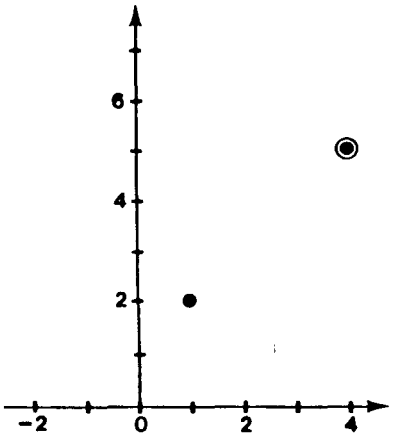


$SS_{xx} = 6$ $\quad \hat{\beta}_1 = 1$ $\quad R^2 = 0{\cdot}43$
$SS_{xy} = 6$ $\quad \hat{\beta}_0 = 1$ $\quad t = 0{\cdot}866$
$SS_{yy} = 14$ $\quad SS_{\hat{y}\hat{y}} = 6$ Signifikanz-
$\bar{x} = 3$ $\quad s^2 = 8$ niveau = 0·273
$\bar{y} = 4$

Beispiel 3: Die *beiden rechten Punkte* aus Beispiel 1 wandern nun gleichmäßig aufeinander zu, bis sie *zusammentreffen*: (1,2), (4,5), (4,5). Wiederum ist es hilfreich, die Studenten nach ihrer Ansicht über die dadurch zu erwartende Veränderung der Kennziffern zu befragen.
Die Regressionsgerade bleibt wieder dieselbe wie in Beispiel 1, ebenso auch die Ausgleichspunkte $\hat{y}_i = \hat{a}+\hat{b}x_i$. Die durch Regression erklärte Streuung, $SS_{\hat{y}\hat{y}}$, bleibt daher gleich.
Alle Punkte liegen nun auf der Regressionsgeraden, die Restvarianz s^2 bzw. SS_e ist Null.
Die Streuung der y-Daten, SS_{yy}, wird kleiner, da alle Punkte nun auf der Regressionsgeraden liegen, ist $SS_{yy} = S_{\hat{y}\hat{y}}$.

Der Anteil der durch Regression erklärten Streuung $r^2 = SS_{\hat{y}\hat{y}}/SS_{yy}$ ist daher 100 %, d.h. $r^2=1$.
Für die t-Statistik folgt aus $s^2=0$, daß sie einen nicht definierten bzw. einen unendlich großen Wert (das würde einen p-Wert von 0 ergeben) hat, obwohl die Steigung der Regressionsgeraden $\hat{b}=1$ ist.

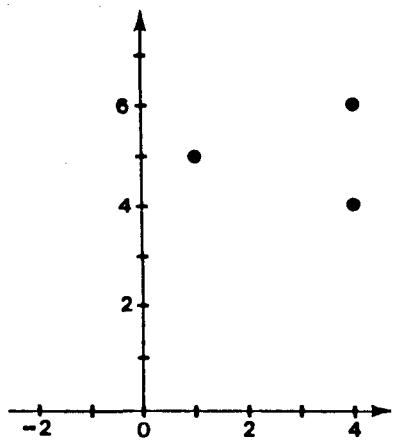
Die angestellten Vergleiche in diesen drei Beispielen klären normalerweise die Unterschiede zwischen der Größe von $\hat{b}$ und der Größe der beobachteten Signifikanz des entsprechenden t-Werts. Eine Aufgabe für die Studenten: Konstruiere eine 3-Punkt-Datenmenge mit $\hat{b}=100$ und einem sehr kleinen t-Wert, der daher eine geringe Signifikanz hat.



$SS_{xx} = 6$ $\quad \hat{\beta}_1 = 1$ $\quad R^2 = 1$
$SS_{xy} = 6$ $\quad \hat{\beta}_0 = 1$ $\quad t = \infty$
$SS_{yy} = 6$ $\quad SS_{\hat{y}\hat{y}} = 6$ $\quad$ Signifikanz-
$\bar{x} = 3$ $\quad s^2 = 0$ $\quad$ niveau = 0
$\bar{y} = 4$

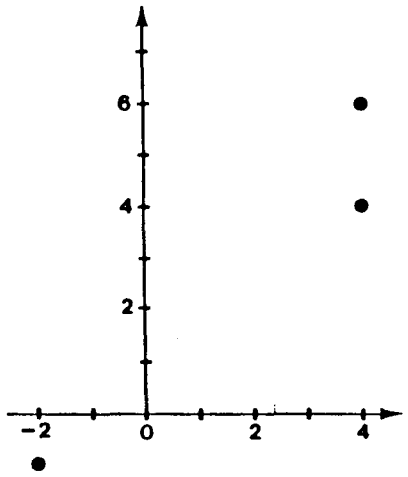
Beispiel 4: Im Vergleich zu Beispiel 1 bewegt sich der *linke Punkt nach oben*: (1,5), (4,4), (4,6). Von den Quadratsummen ändert sich nun SS_{xy} und S_{yy}, daraus ergibt sich $\hat{b}=0$ und $\hat{a}=5$. Dadurch wird klar, welche Rolle dieser einzelne Punkt für Achsenabschnitt und Steigung der Regressionsgeraden spielt.
Die Regressionsgerade ist parallel zur 1. Achse, die Ausgleichspunkte $\hat{y}_i$ streuen überhaupt nicht, sie sind alle identisch, d.h. $SS_{\hat{y}\hat{y}}=0$. Die durch diese Regressionsgerade erklärte Streuung der y-Daten um $\bar{y}$ ist Null, daher ist $r^2=0$. In so einem Fall kann man ein y_i geradesogut durch den Mittelwert $\bar{y}$ schätzen wie durch den Ausgleichspunkt $\hat{y}_i = \hat{a}+\hat{b}x_i$. Darin liegt nach Ansicht des Autors eindeutig der Reiz des Beispiels für die Studenten. Darüberhinaus

sollte man beachten, daß die Schätzung von σ^2 unverändert zu Beispiel 1 ist.



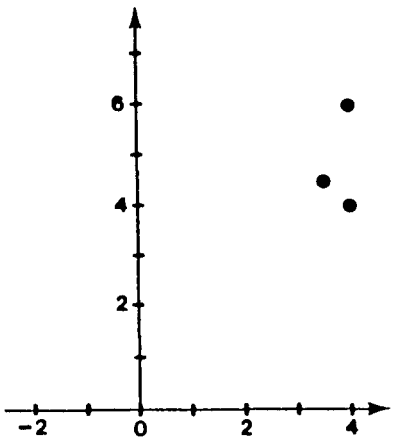
$SS_{xx} = 6$	$\hat{\beta}_1 = 0$	$R^2 = 0$
$SS_{xy} = 0$	$\hat{\beta}_0 = 5$	$t = 0$
$SS_{yy} = 2$	$SS_{\hat{y}\hat{y}} = 0$	Signifikanz-
$\bar{x} = 3$	$s^2 = 2$	niveau $= 0{\cdot}50$
$\bar{y} = 5$		

Beispiel 5: Der *linke Punkt* in Beispiel 1 wird längs der ursprünglichen Regressionsgeraden weiter *nach links außen bewegt*: (-2,-1), (4,4), (4,6). Dadurch werden die Quadratsummen SS_{xx}, SS_{yy} und SS_{xy} größer. Man beachte aber, daß die Koeffizienten $\hat{a}$ und $\hat{b}$ sowie s^2 unverändert bleiben. Das Hinauswandern des linken Punktes hat auch einen kleineren Ausgleichswert $\hat{y}$ zur Folge, auch $S_{\hat{y}\hat{y}}$ wird damit größer, jedoch bezogen auf S_{yy} in stärkerem Ausmaß, wovon man sich durch Rechnung überzeugen kann.
Das bedeutet, daß r^2 größer wird. Leichter einzusehen ist, daß der Wert der t-Statistik größer wird, d.h. sein p-Wert wird kleiner, seine Signifikanz größer.
Dies macht insgesamt deutlich, daß Punkte, die sich weiter außen befinden, die Regressionsgerade stabilisieren und der Schätzung von b eine größere Sicherheit verleihen (daß der Wert b von Null wirklich verschieden ist), auch für den Fall, daß die Schätzung von σ^2 sich nicht verändert hat.



$SS_{xx} = 24$ $\quad \hat{\beta}_1 = 1$ $\quad R^2 = 0{\cdot}923$
$SS_{xy} = 24$ $\quad \hat{\beta}_0 = 1$ $\quad t = 3{\cdot}46$
$SS_{yy} = 26$ $\quad SS_{\hat{y}\hat{y}} = 24$ Signifikanz-
$\bar{x} = 2$ $\quad s^2 = 2$ niveau = 0·090
$\bar{y} = 3$

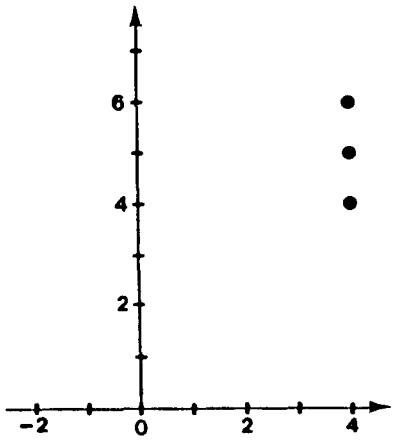
Beispiel 6: Der *linke Punkt* in Beispiel 1 wird nun längs der alten Regressionsgeraden *näher* zu den beiden anderen Punkten gerückt: (3,5, 4,5), (4,4), (4,6). Obwohl alle Quadratsummen SS_{xx}, SS_{yy} und SS_{xy} sich verändern, bleibt die Ausgleichsgerade dieselbe wie in Beispiel 1, ebenso bleibt auch s^2 gleich.
Die Rechnung ergibt, daß r^2 sowie der beobachtete t-Wert für $\hat{b}$ kleiner werden (der p-Wert von $\hat{b}$ steigt).
Die Änderung im Vergleich zu Beispiel 1 ist genau gegengleich wie in Beispiel 5.



$SS_{xx} = \frac{1}{6}$ $\quad \hat{\beta}_1 = 1$ $\quad R^2 = 0{\cdot}0769$
$SS_{xy} = \frac{1}{6}$ $\quad \hat{\beta}_0 = 1$ $\quad t = 0{\cdot}289$
$SS_{yy} = 2\frac{1}{6}$ $\quad SS_{\hat{y}\hat{y}} = \frac{1}{6}$ Signifikanz-
$\bar{x} = \frac{23}{6}$ $\quad s^2 = 2$ niveau = 0·410
$\bar{y} = \frac{29}{6}$

Beispiel 7: Dieses Beispiel ist nur für Klassen mit überdurchschnittlichem mathematischen Hintergrund gedacht. Der linke Punkt bewegt sich wie in Beispiel 6, jedoch noch näher an die beiden anderen Punkte heran, bis sie *alle drei auf einer senkrechten Ge-*

raden "aufgefädelt" sind: (4,5), (4,4), (4,6). In diesem Fall verhindert $SS_{xx}=0$ eine Anwendung der üblichen Formeln. Eine neuerliche Herleitung der Normalgleichungen zeigt nun, da $x_1=x_2=x_3=4$, daß wir nur *eine* linear unabhängige Gleichung in zwei Unbekannten $\hat{a}$ und $\hat{b}$ haben. Wir können daher eine Variable frei wählen, die andere ergibt sich dann. Die Wahl verschiedener Werte von $\hat{b}$, z.B. $\hat{b}=1,0$ und -1 und das Einzeichnen der Ausgleichsgeraden gibt den Studenten ein Gefühl für solche Fälle.



$SS_{xx} = 0$ $\quad \hat{\beta}_1 = \text{undefiniert}$ $\quad R^2 = \text{undefiniert}$

$SS_{xy} = 0$ $\quad \hat{\beta}_0 = \text{undefiniert}$ $\quad t = \text{undefiniert}$

$SS_{yy} = 2$ $\quad SS_{\hat{y}\hat{y}} = \text{undefiniert}$

$\bar{x} = 4$ $\quad s^2 = \text{undefiniert}$

$\bar{y} = 5$

3. Zusammenfassung und Schlußfolgerung

Regression ist eine in der Praxis weit verbreitete und überaus nützliche statistische Methode. Dennoch passiert es häufig, daß die Konzepte für den Lernenden von den Daten verdunkelt werden. Durch Reduzieren der Datenmenge und durch systematische Veränderung eines Punktes oder zweier symmetrisch liegender Punkte können die Berechnungen minimiert werden. Gerade dadurch wird üblicherweise die Einsicht unterstützt.

Literatur:

Ott, Lymann: In Introduction to Statistical Methods and Data Analysis. Duxbury, 1984.