Bei Vorhersagen sind manche Ausgangsdaten unwichtig[1]

TERESA L. BITTNER, USA

[1] Das Original erschien in Teaching Statistics (Volume 35, Number 2, Summer 2013; S. 80–83). Originaltitel: A limitation with least squares predictions
Übersetzung, Bearbeitung und Kürzung: J. MEYER

Zusammenfassung: *Manche Vorhersagen werden mit Hilfe von Regressionsgeraden vorgenommen; die mit dem Algorithmus der kleinsten Quadrate ermittelt werden. Es ist schon häufiger festgestellt worden, dass manche Ausgangsdaten einen größeren Einfluss auf die Vorhersage haben als andere, aber es ist fast unbekannt, dass manche Ausgangsdaten gar keinen Einfluss auf die Vorhersage haben.*

1 Einleitung

Zu n Datenpunkten $(x_i; y_i)$ soll eine Regressionsgerade ermittelt werden. Man modelliert also mit dem Term $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$, wobei die ε_i Fehlerterme darstellen, deren Quadratsumme minimiert werden soll. Wenn die angepassten Geradenparameter mit $\hat{\beta}_0$ und $\hat{\beta}_1$ bezeichnet werden, gilt $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$ sowie

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2};$$ dabei ist $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ und

$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$. Damit ist $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$ die Schätzung für den zu x gehörigen y-Wert.

Worum es bei diesem Artikel geht, wird am besten durch ein einfaches Beispiel deutlich. Die Anmeldezahlen für einen Kindergarten in den Jahren 2001 bis 2004 sind durch die geordneten Zahlenpaare (1; 62), (2; 43), (3; 78) und (4; 82) gegeben. Sie werden in Abb. 1 zusammen mit der Ausgleichsgerade (durchgezogen) dargestellt.

Die Ausgleichsgerade hat die Gleichung $\hat{y} = 42{,}5 + 9{,}5 \cdot x$. Um damit die Anzahl der Anmeldungen für 2005 vorherzusagen, wird $x = 5$ eingesetzt, was zu $\hat{y} = 90$ führt.

Nun nehme man an, dass bei der Datenerhebung ein Fehler passiert sei und dass 2002 nur 20 Kinder angemeldet wurden. Dies führt zu $\hat{y} = 31 + 11{,}8 \cdot x$ und damit überraschenderweise wiederum zu der Vorhersage $\hat{y} = 90$ für das Jahr 2005.

Sind stattdessen im Jahr 2002 sogar 120 Kinder angemeldet worden, führt auch das zur Vorhersage $\hat{y} = 81 + 1{,}8 \cdot 5 = 90$.

Ganz unterschiedliche Werte für y_2 haben zur selben Vorhersage geführt. In Abb. 1 werden die drei verschiedenen Regressionsgeraden zusammen dargestellt.

Im Folgenden wird dies Phänomen algebraisch untersucht.

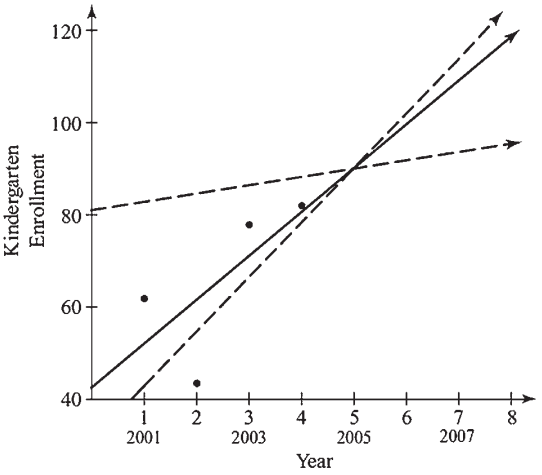

Abb. 1: Drei Regressionsgeraden und die ursprünglichen Datenpunkte

2 Punkte, die nicht zur Vorhersage beitragen

In Fällen mit teurer oder sehr schwieriger Datenermittlung ist es wichtig zu wissen, ob man Datenpunkte auch einfach weglassen kann, ohne die Qualität der Vorhersage zu beeinträchtigen. Bisher ist kaum untersucht worden, ob Datenpunkte nur einen minimalen oder gar keinen Einfluss auf die Prognose haben.

Die Frage lautet also: Wie ändert sich die Vorhersage eines Modells, wenn die Eingangsgrößen sich ändern? Im vorliegenden Fall heißt das:

Die zu den Datenpunkten $(x_1; y_1)$, …, $(x_n; y_n)$ gehörige Regressionsgerade hat die Gleichung $\hat{y} = \bar{y} + \hat{\beta}_1 \cdot (x - \bar{x})$. Wir wollen den zu $x = x_p$ gehörigen y-Wert voraussagen, der Voraussagewert sei mit $\hat{y}_p$ bezeichnet. Dann ist

$$\hat{y}_p = \bar{y} + \hat{\beta}_1 \cdot (x_p - \bar{x}). \qquad (1)$$

Nun wird der Datenpunkt $(x_d; y_d)$ geändert zu $(x_d; y_d + \Delta)$, und es sei $\tilde{y}_p$ der neue zu x_p gehörige Vorhersagewert.

Aus (1) bekommt man

$$\tilde{y}_p = \left(\bar{y} + \frac{\Delta}{n}\right) + \left(\hat{\beta}_1 + \frac{(x_d - \bar{x}) \cdot \Delta}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \tag{2}$$

$$= \hat{y}_p + \Delta \cdot \left(\frac{1}{n} + \frac{(x_d - \bar{x}) \cdot (x_p - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Ist Δ klein, werden sich die beiden Vorhersagen $\hat{y}_p$ und $\tilde{y}_p$ kaum unterscheiden. Der Unterschied verschwindet (auch für $\Delta \neq 0$), wenn

$$(x_d - \bar{x}) \cdot (x_p - \bar{x}) = -\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3}$$

bzw. wenn

$$x_p = \bar{x} + \frac{1}{n} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x} - x_d} \tag{4}$$

gilt. Falls (4) gilt, so ist die zu x_p gehörige Vorhersage von x_d unabhängig.

Dies wird durch das Kindergarten-Beispiel bestätigt: Es ist

$$x_p = \frac{10}{4} + \frac{1}{4} \cdot \frac{(-1{,}5)^2 + (-0{,}5)^2 + 0{,}5^2 + 1{,}5^2}{2{,}5 - 2} = 5;$$

d. h. die Vorhersage für 2005 ist vom zu 2002 gehörigen Datenpunkt unabhängig.

3 Wie vermeidet man unwichtige Datenpunkte?

Es wird hilfreich sein, (4) nach x_d aufzulösen; man bekommt

$$x_d = \frac{\sum_{i=1}^{n} x_i^2 - x_p \cdot \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i - x_p \cdot n}. \tag{5}$$

Gleichung (5) kann allerdings unbrauchbare x-Werte liefern; für das Kindergarten-Beispiel könnten nichtganze x-Werte herauskommen. Bekommt man etwa $x_d = 1{,}01$, so wird man sagen können, dass der Datenpunkt (1; …) nur ganz schwach zur Vorhersage beiträgt.

Es folgt ein einfaches Beispiel, in dem die Ausgaben der US-amerikanischen Regierung vorhergesagt werden sollen. Grundlage sind die Datenpunkte

(1999; 2761,9), (2001; 3093,6), (2003; 3458,6) (2005; 3916,4), (2007; 4430,0), (2009; 4998,8).

Mit Hilfe von (5) werden die Jahre identifiziert, die für die Vorhersage keinen Einfluss haben; das Ergebnis ist:

Für 2010 hat 2002,06 keinen Einfluss;
für 2011 hat 2002,33 keinen Einfluss;
für 2012 hat 2002,54 keinen Einfluss;
für 2013 hat 2002,70 keinen Einfluss;
für 2014 hat 2002,83 keinen Einfluss.

Für 2010 hat 2002 fast keinen Einfluss, aber je weiter man in die Zukunft vorschreitet, umso größer wird der Einfluss von 2002; er bleibt gleichwohl insgesamt relativ klein. Würde man jedem Datenpunkt die Kosten seiner Erhebung zuordnen, könnte man den Befund dazu benutzen, um zu diskutieren, ob überhaupt Daten für 2002 erhoben werden sollten.

Literatur

Bittner, T.; Norlin, K. (2006): The noncontribution of some data in least squares regression predictions. In: *International Journal of Applied Management and Technology* 4 (1), S. 231–244.

Anschrift der Verfasserin

Teresa L. Bittner
Walden University, Minneapolis, Minnesota, USA