EVALUATIONEN UND IHRE EINFLUSSFAKTOREN

Werden Studenten altersmilde?

>> **von Wolfgang Gohout >** Seit 2006 finden gemäß Evaluationsordnung an der Hochschule Pforzheim stichprobenartig Evaluationen von Lehrveranstaltungen statt. Als Vorsitzender der Studienkommission Wirtschaftsingenieurwesen koordiniere ich – gemeinsam mit der Studienkommission – diese Evaluationen im Bereich Wirtschaftsingenieurwesen. So konnte ich ein paar Eckdaten beobachten und hatte einen Anfangsverdacht: je kleiner die Gruppe und je höher das Fachsemester, desto besser die Evaluationsnote. Dies ist sicher keine besonders überraschende Beobachtung. Dennoch wollte ich sie etwas genauer untersuchen. Dabei sollen der Einfachheit halber das Fachsemester, die Gruppengröße und die Evaluationsnote als kardinale Merkmale aufgefasst werden.

In 65 evaluierten Lehrveranstaltungen des Bereichs Wirtschaftsingenieurwesen haben sich zwischen der Semesterzahl und der Evaluationsnote ein Korrelationskoeffizient von - 47% und zwischen der Gruppengröße und der Evaluationsnote ein Korrelationskoeffizient von 42% ergeben. Die Vorzeichen entsprechen der vermuteten Richtung des jeweiligen Zusammenhangs. Die Beträge von mehr als 40% sprechen für eine durchaus nennenswerte Stärke des Zusammenhangs.
Eine lineare Einfachregression der Evaluationsnote auf das Semester liefert die Regressionsgerade:

$$\text{Evaluationsnote} = 2{,}2257 - 0{,}0875 \cdot \text{Semester}$$

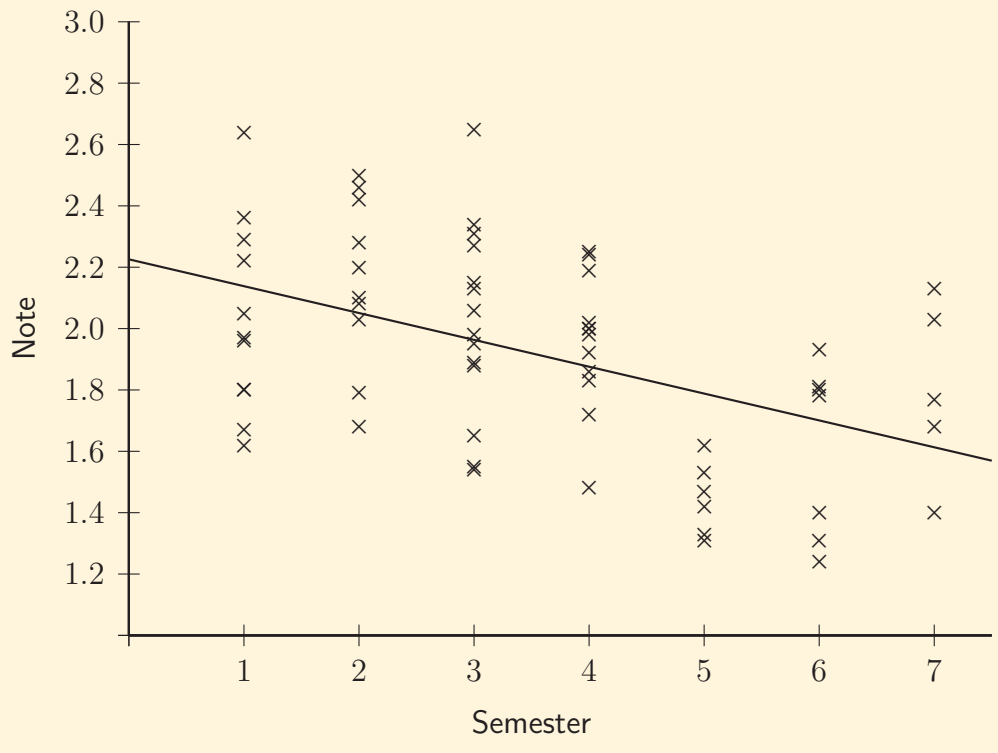

Abbildung 1: Streuungsdiagramm und Regressionsgerade der Regression auf das Semester.

Abbildung 1 zeigt das Streuungsdiagramm mit der Regressionsgerade. Mit jedem zusätzlichen Fachsemester sinkt die Evaluationsnote um durchschnittlich 0,0875. Die Residuen sind nach dem Schiefe-Kurtosis-Test von Jarque/Bera normalverteilt, so dass die t-Werte der geschätzten Parameter der Regressionsgeraden zuverlässig interpretiert werden können. Da ihr Betrag größer als 4 ist, ist die Regressionsschätzung hoch signifikant. Werden die Studenten etwa altersmilde?

Die Einfachregression der Evaluationsnote auf die Gruppengröße liefert die Regressionsgerade:

$$\text{Evaluationsnote} = 1{,}6901 + 0{,}0070 \cdot \text{Gruppengröße}$$

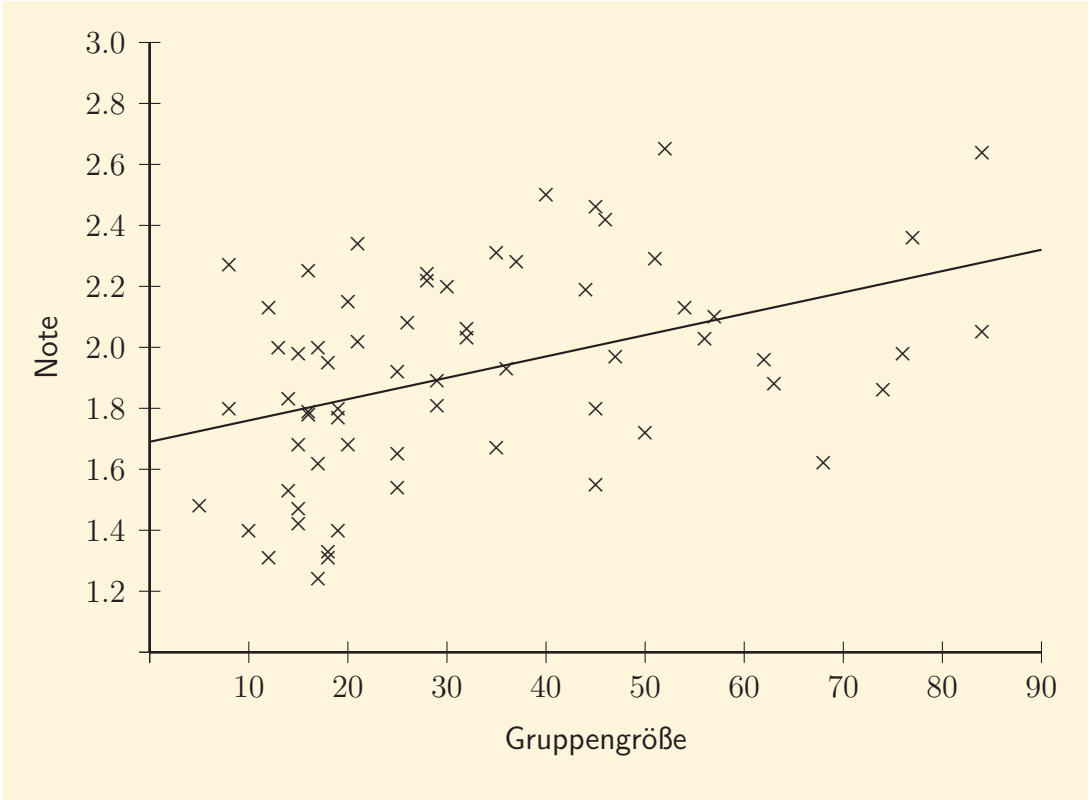

Abbildung 2: Streuungsdiagramm und Regressionsgerade der Regression auf die Gruppengröße.

Abbildung 2 zeigt das zugehörige Streuungsdiagramm mit der Regressionsgerade. Mit jedem zusätzlichen Studenten steigt die Evaluationsnote um durchschnittlich 0,007. Die Residuen sind wieder normalverteilt. Die t-Werte sind abermals hoch signifikant. Diese Wirkungsrichtung scheint durchaus plausibel. In kleineren Gruppen ist der Dozent besser zu verstehen und der Kontakt zu den Studierenden ist enger.

Nun gibt es aber auch einen Zusammenhang zwischen den beiden Regressoren selbst. Die Korrelation zwischen dem Fachsemester und der Gruppengröße beträgt - 57%. In den höheren Semestern sind die Gruppen eben kleiner als in den unteren Semestern. Eine lineare Zweifachregression der Evaluationsnote auf beide Regressoren führt wegen dieser Korrelation zwar nicht mehr zu zuverlässig schätzbaren Koeffizienten, aber zu einer recht guten Erklärungs- und Prognosegüte der Evaluationsnote:

$$\text{Evaluationsnote} = 2{,}0192 - 0{,}0636 \cdot \text{Semester} + 0{,}0038 \cdot \text{Gruppengröße}$$

Das korrigierte Bestimmtheitsmaß dieser Regression beträgt etwa 25%, was für einen echten Erklärungsversuch vielleicht nicht viel sein mag, aber sicherlich zu viel ist, wenn man bedenkt, dass die Evaluationsnote ja nicht wirklich durch diese extrinsischen Faktoren erklärt werden soll, sondern durch die 14 intrinsischen Faktoren beziehungsweise Items des Evaluationsfragebogens.

Wenn man die Evaluationsnote von dem Einfluss der extrinsischen Faktoren bereinigen möchte und von der künftigen Gültigkeit des geschätzten Zusammenhangs ausgeht, dann müsste man die Note einer Evaluation folgendermaßen korrigieren:

$$\text{neue Note} = \text{Note} + \text{MW(Note)} - 2{,}0192 + 0{,}0636 \cdot \text{Semester} - 0{,}0038 \cdot \text{Gruppengröße}$$

wobei MW(Note) = 1,9188 der Mittelwert aller 65 Evaluationsnoten ist.

Diese Korrektur sollte dann vorgenommen werden, wenn man Evaluationsnoten verschiedener Lehrveranstaltungen fair miteinander vergleichen möchte, was aber sicher nicht das primäre Ziel der Evaluation ist. Aus der Note 3,5 im 1. Semester mit 75 Teilnehmern würde dann die „faire" Note 3,18. Und aus der Note 1,4 im 7. Semester mit 15 Teilnehmern würde die korrigierte Note 1,69.

>

Eine etwas andere Problematik besteht in der Korrelation der intrinsischen Faktoren selbst. Die Items oder Aussagen des Evaluationsfragebogens, die jeweils mit 1, 2, 3, 4, oder 5 zu bewerten sind, lauten:

1. Das Lernziel der Lehrveranstaltung war von Beginn an klar ersichtlich und wurde konsequent verfolgt.
2. Der/die Lehrende stellt einen guten Bezug zur Praxis her.
3. … wirkt gut vorbereitet.
4. … erklärt schwierige Sachverhalte gut.
5. … regt zum eigenständigen Denken an.
6. … ist gut zu verstehen (Lautstärke, Sprechtempo, Ausdrucksweise).
7. … behandelt Studierende fair.
8. … tritt sicher auf.
9. … zeigt Engagement in seiner/ihrer Lehrtätigkeit.
10. … weckt Interesse am Thema der Veranstaltung.
11. … schafft eine gute Arbeitsatmosphäre.
12. … setzt Medien … angemessen und hilfreich ein …
13. Ich habe mich aktiv an der Veranstaltung beteiligt.
14. Ich war von vorneherein besonders an dieser Veranstaltung interessiert.

Die Einzelnoten für diese 14 Punkte zeigen im Bereich Wirtschaftsingenieurwesen (für $n = 60$ Datensätze) folgende Auffälligkeit: Die Fragen 2, 4, 10 und 11 weisen paarweise Korrelationskoeffizienten zwischen 78% und 91% auf und sind damit extrem multikollinear. Aber auch alle 14 Items zusammen sind extrem multikollinear: die Determinante der Korrelationsmatrix hat mit etwa 10^{-6} einen Wert von praktisch Null; und der Bartlett-Sphärentest auf fehlende Multikollinearität hat einen p-Wert von etwa 10^{-100}. Die Multikollinearität ist also gigantisch.

Um einen (quantitativen) Eindruck von der „Güte" einer Lehrveranstaltung zu bekommen, könnte man multikollineare, also redundante Items durchaus auf ein „repräsentatives" Item oder einen gemeinsamen Faktor reduzieren. Eine Faktorenanalyse zeigt ebenfalls sehr eindeutig, dass es für alle 14 Items nur einen gemeinsamen Faktor gibt. Andererseits kann es bei einer einzelnen Evaluation für die Dozentin oder den Dozenten durchaus wichtige Hinweise auf Verbesserungspotential geben, wenn man alle Items beibehält. Lediglich die Items 13 und 14 scheinen – nicht nur mir – nicht auf die Güte der Veranstaltung zu zielen, sondern eher auf die Einstellung der Studierenden. Auch die Abbildung 3 zeigt, dass sie in der Darstellung der Faktorladungen der Items in einem Zwei-Faktoren-Modell eine äußere Position einnehmen. Ansonsten zeigt die Abbildung 3, dass sich alle 14 Items viel stärker um die horizontale Achse des ersten Faktors gruppieren als um die vertikale Achse eines möglichen zweiten Faktors. Ein zweiter Faktor ist daher irrelevant.

Aber auch wenn die multikollinearen Items den einzelnen Lehrenden individuelle Verbesserungspotenziale aufzeigen können, sollte man sich des Effektes korrelierter Aussagen bewusst sein. Um diesen Effekt zu illustrieren, soll hier ein extremer Fall betrachtet werden: Nehmen wir an, Item 1 ist unkorreliert von den restlichen (n –1) Items, die ihrerseits paarweise perfekt korreliert sind mit dem Korrelationskoeffizienten 1. Die Streuung soll für alle Items gleich groß sein. Man kann zeigen, dass dann der Korrelationskoeffizient r zwischen dem Mittelwert aller n Noten und jeder einzelnen Note der perfekt korrelierten Items 2 bis n der folgenden Gleichung genügt:

$$r = \frac{n-1}{\sqrt{1+(n-1)^2}}$$

Dieser Ausdruck geht mit wachsendem n schnell gegen 1. Für $n = 5$ also bei vier perfekt korrelierten Items beträgt die Korrelation beispielsweise 97%. Dies bedeutet, dass man bei der Gestaltung des Fragebogens eine bestimmte Eigenschaft einer Dozentin oder eines Dozenten allein dadurch (beliebig) stärker gewichten kann, dass man die entsprechende Aussage quasi identisch, aber mit anderen Worten mehrfach wiederholt, z.B. „… ist sympathisch", „… ist nett", „… ist fair" und so weiter. Zusammenfassend kann man festhalten, dass das Fachsemester und die Gruppengröße einen nicht unerheblichen Einfluss auf die Evaluationsnote haben – was aber vermutlich nicht auf „Altersmilde" der Studierenden zurückzuführen ist. Andererseits sollte man sich bei der Konzeption eines Fragebogens stets des wechselseitigen Einflusses der Items auf die „Gesamtnote" bewusst sein. Im Idealfall sollten die Items daher unabhängig sein, wenn man nur an der (eindimensionalen) Bewertung der Veranstaltungen interessiert ist. Wenn man aber individuelle Schlussfolgerungen zulassen will, dann sollte man zumindest einige der redundanten Items beibehalten. Die Items 13 und 14 könnten meines Erachtens gestrichen werden, da sie sich nicht auf die Bewertung der Veranstaltung beziehen. Dagegen wäre eine Frage nach der persönlichen „workload" im Sinn von durchschnittlich aufgewendeten Semesterwochenstunden für die zu evaluierende Veranstaltung sinnvoll und hilfreich. Dieses Instrument der „workload-Messung" ist von großer Bedeutung im Hinblick auf Akkreditierungen und scheint mir bislang weitgehend zu fehlen. In die Evaluationsnote dürfte und könnte diese Zusatzfrage natürlich nicht eingehen, da sie ja nicht der 5-Punkt-Likert-Skala entspricht

Dr. Dr. habil. Wolfgang Gohout
ist Professor für Quantitative Methoden im Bereich Wirtschaftsingenieurwesen.

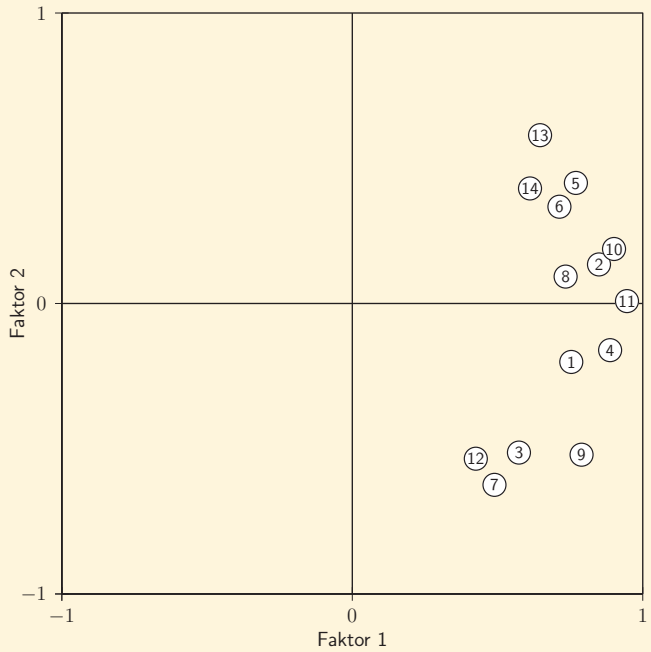

Abbildung 3: Ladungen der 14 Items in der Zwei-Faktoren-Ebene.