

## Das Erkennen von Multikollinearität mittels der PS-Explore-Korrelationsanalyse

Von Multikollinearität spricht man speziell bei der Regressionsanalyse, wenn zwei oder mehrere Regressoren (Einflussgrößen) hoch miteinander korreliert sind.

Im Falle perfekter Kollinearität ist die rechnerische Durchführung der linearen Regressionsanalyse theoretisch unmöglich. Wenigstens eine der Einflussgrößen lässt sich dann als Linearkombination von einem oder mehreren der anderen Regressoren darstellen.

Mathematisch lässt sich die Lösung des linearen Regressionsproblems

$$y_i = b_0 + b_1x_{i,1} + \dots + b_px_{i,p}$$

für die Regressionskoeffizienten mit der Kleinste-Quadrate-Methode darstellen als

$$\hat{b} = (X'X)^{-1}X'y$$

Bei vollständiger Kollinearität ist die Matrix  $X'X$  singulär und kann deshalb nicht invertiert werden. Methodisch stellt daher Kollinearität eigentlich kein Problem dar und ist sehr einfach zu erkennen, da der Schätzer  $b$  gar nicht berechnet werden kann.

*Vollständige Kollinearität kann sehr leicht als „dummer“ Fehler auftreten, wenn man unsachgemäß mit so genannten Dummy-Variablen arbeitet.* Die Dummy-Kodierung wird etwa benutzt, wenn man qualitative Merkmale in der eigentlich auf quantitative Einflussgrößen abgestellten Regressionsanalyse benutzt. Hierbei werden die  $k$  Klassen eines qualitativen Merkmals in  $k$  Einzelmerkmale aufgelöst, die nur Wertausprägungen von 0 und 1 besitzen. I.d.R. 0 für „Klassenzugehörigkeit liegt nicht vor“ und 1 für „Klassenzugehörigkeit liegt vor“. Im Falle etwa des Merkmals „Geschlecht“ könnten 2 Merkmale generiert werden: Merkmal „weiblich“ und Merkmal „männlich“. Würde man nun beide Merkmale in die Regressionsanalyse übernehmen, dann bestünde Singularität im o.g. Sinne und das Gleichungssystem wäre nicht lösbar. Im Falle des Geschlechts könnte es aber reichen nur ein Dummy-Merkmal in das System aufzunehmen und die Sache wäre methodisch sauber. Schwieriger wird es, wenn man für jedes Dummy bzw. jede Klasse des ursprünglichen Merkmals einen eigenen Gewichtungsfaktor haben will. In einem solchen Fall empfiehlt sich dann eine Kovarianzanalyse, die unter Einführung einer hier nicht näher erörterten Zusatzannahme dann zu einem sauberen Ergebnis führt.

Es kann aber selbst bei Fällen der unsachgemäßen Regressionsanalyse mit Dummies passieren, dass scheinbar eine Lösung gefunden wird. Dies hängt zusammen mit der numerischen Genauigkeit (besser „Ungenauigkeit“) eines Computers. Da mit begrenzter Genauigkeit gerechnet wird kann es sein, dass die Lösung des Gleichungssystems immer „haarscharf“ an der Singularität vorbeischliddert.

Über dieses Vorbeischliddern sollte man nicht leichtfertig hinwegsehen. Das Verfahren zum Schätzen der Regressionskoeffizienten wird mit zunehmender Multikollinearität instabil und die berechneten Regressionskoeffizienten sind nicht mehr sehr präzise. Es ist dann auch möglich bzw. zu erwarten, dass nur leichte Änderungen in der Datenstichprobe zu ganz anderen Regressionskoeffizienten führen. Sogar das Vorzeichen der Koeffizienten kann sich ändern.

Die Identifikation der Multikollinearität kann mittels verschiedener Maße aufgedeckt werden. Die bekanntesten dürften die *Toleranz*

$$T_i = 1 - R_i^2$$

und der *Varianzinflationsfaktor* sein:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Im Falle der Toleranz geht man davon aus, dass bei einem Wert  $< 0.2$  Multikollinearität vorliegen kann. Beim Varianzinflationsfaktor deutet ein Wert  $> 5$  auf Kollinearität hin.

$R_i$  steht oben für das Bestimmtheitsmaß, wobei eine Regression aller anderen Einflussgrößen  $j \neq i$  auf jeweils eine verbleibende Einflussgröße  $i$  gerechnet wird. Ist das Bestimmtheitsmaß dann sehr hoch, so ist klar, dass Merkmal  $i$  sehr gut durch die anderen Merkmale vorhergesagt werden kann und also (näherungsweise) Multikollinearität vorliegt.

In der neuen ad-hoc-Korrelationsanalyse von PS-Explore lässt sich der Varianzinflationsfaktor sehr leicht berechnen. In der nachstehenden Abbildung sieht man das Button VIF. Man klickt dieses an und erhält dann automatisch in der Hauptdiagonalen der berechneten Korrelationsmatrix den zu jedem Merkmal gehörigen VIF. Werte über 5 werden durch Ausrufezeichen besonders hervorgehoben:

Ad-Hoc-Korrelationen

sig. 0,05

VIF

	Blüte.B	Blüte.L	Kelch.B	Kelch.L
Blüte.B	16,09!!!	0,96	-0,37	0,82
Blüte.L	0,96	31,26!!!	-0,43	0,87
Kelch.B	-0,37	-0,43	2,10	-0,12
Kelch.L	0,82	0,87	-0,12	7,07!!!

Art: Varianzinflationsfaktor

Achtung: Für Merkmal Blüte.L liegt der Varianzinflationsfaktor über 5.  
Dies ist ein Hinweis auf starke Multikollinearität. Es wird empfohlen das Merkmal, aus den Berechnungen zu entfernen.

Blüte.B  
Blüte.L  
Kelch.B  
Kelch.L

Identifikationsmerkmal:  
lfd Nr

Berechnung mit optionaler Zielgröße:  
-

Berechnung mit optionalem Kontrollmerkmal:  
-

berechnen beenden

Entfernt man sodann die im Hinweisfenster angegebene Größe (hier Blüte.L) aus der Berechnung, reguliert sich die Situation und die VIF der verbleibenden Regressanden bleiben im regulären Bereich.

Im abschließenden Beispiel wurde nun absichtlich der oben beschriebene „dumme Fehler“ bei der Dummy-Kodierung begangen. Man sieht sofort, dass PS-Explore diesen Fehler gnadenlos bloßstellt und die Herausnahme wenigstens eines Dummies empfiehlt:

	norm.Kaufpreis	Nettokaltniete	Ursprungsbaujahr	Wohnlage_einfach	Wohnlage_mittel	Wohnlage_gut
norm.Kaufpreis	1,59	0,64	0,57	-0,25	-0,14	0,40
Nettokaltniete	0,64	3,86	0,82	-0,40	-0,10	0,50
Ursprungsbaujahr	0,57	0,82	1,71	-0,40	0,01	0,37
Wohnlage_einfach	-0,25	-0,40	-0,40	Extreme Multikollinearität	0,58 \$	-0,26
Wohnlage_mittel				0,58 \$	Extreme Multikollinearität	0,63 \$
Wohnlage_gut				-0,26	0,63 \$	Extreme Multikollinearität

Achtung: Für Merkmal Wohnlage mittel liegt der Varianzinflationsfaktor über 5.  
Dies ist ein Hinweis auf starke Multikollinearität. Es wird empfohlen das Merkmal, aus den Berechnungen zu entfernen.

Hier bliebe nun der Ausweg über die Kovarianzanalyse um für alle drei Wohnlagen des Beispiels dennoch valide Koeffizienten zu berechnen. Eine Dummy-Kodierung braucht dabei nicht vorgenommen zu werden. Die PS-Explore Regressions- und Kovarianzanalyse erkennt bekanntlich sofort automatisch, wann und wie welche Analyseform zu wählen ist.

Wer sich näher über das Problem der Multikollinearität informieren möchte, der sei hier einfach auf zwei im Internet verfügbare Artikel verwiesen:

- 1) <http://de.wikipedia.org/wiki/Multikollinearit%C3%A4t>
- 2) <http://www.uibk.ac.at/econometrics/einf/07p.pdf>

Artikel 1 beschreibt das Phänomen kurz und bündig während Artikel 2 sehr gut und sehr ausführlich auf das Problem eingeht.