Regressionsanalyse (2)

Wilfried Mann,
Mettmann

Die multiple lineare Regressionsanalyse ist in der Praxis ein iterativer Prozess, welcher einen geeigneten, zufälligen Datensatz und das schrittweise Zulassen und Weglassen von Variablen eines sachverständig gewählten, statistischen Modells erfordert.

1. Ziele

Zunächst müssen die Ziele der Datenanalyse festgelegt werden. Geht es darum, Erkenntnisse zu grundsätzlichen Zusammenhängen in einem Gutachterausschussgebiet, Bundesland oder der Bundesrepublik zu gewinnen, werden sehr viele Kauffälle benötigt, um robuste Ergebnisse zu generieren.

Der varianzanalytische Teil der Analyse gibt dann Hinweise auf signifikante Einflussgrößen im Modell und beantwortet eine erste Frage in Bezug auf das Datenmaterial:

Welche Einflussmerkmale bestimmen die Zielgröße?

Die nach dem Ausgleichungsprinzip ermittelte Regressionsgleichung zeigt die funktionalen Zusammenhänge zwischen den Einflussgrößen und der Zielgröße auf. Hieraus lassen sich Werte für Normobjekte (z.B. Bodenwerte, Immobilienrichtwerte, Vergleichswerte) ableiten. Eine weitere Möglichkeit besteht in der Ableitung von Umrechnungskoeffizienten und Indexreihen und beantwortet somit eine zweite Frage:

Wie groß sind die Unterschiede in den gruppierten Ausprägungen der Einflussgrößen oder wie stellen sich funktionale Zusammenhänge dar?

Es kann auch das Ziel sein, für einen kleinräumigen Bereich Aussagen zu einem Vergleichswert zu treffen. Qualitätsgrenzen der Analyse werden dann allerdings durch die geringen Fallzahlen gesetzt.

2. Schritte im Prozess

2.1 Datenaufbereitung

In einem zuvor definierten Untersuchungsgebiet, z.B. Einfamilienhäuser in einem Gutachterausschussgebiet, ist die Datenmenge letztlich abhängig von der Anzahl der unabhängigen Einflussgrößen, die im Regressionsmodell festgelegt sind (Empfehlung n > 1.000). Reichen die Fälle aus einem Jahr nicht aus, dann können Vorjahre berücksichtigt werden (Zeitreihenanalyse).

Zielgröße

Die Zielgröße ist eindeutig zu definieren. Dies können z.B. der bereinigte Kaufpreis (unter Berücksichtigung von besonderen objektspezifischen Grundstücksmerkmalen) pro m^2 Grundstücksfläche bzw. Wohnfläche, der Liegenschaftszinssatz oder der Sachwertfaktor sein.

Unabhängige Variable

Die Auswahl erfolgt in Verbindung mit der Zielgröße zunächst sachverständig. Es ist darauf zu achten, dass gruppierte Variablen (pro Ausprägung mindestens 10 Fälle) vollständig und möglichst gleichverteilt, und verhältnisskalierte Variablen vollständig und möglichst normalverteilt vorliegen. Untypische Merkmale, wie z.B. Baujahr < 1.850, sind auszugrenzen. Eventuelle Korrelationen zwischen zwei Merkmalen können sachlogisch oder durch einen Vorab-Korrelationstest festgestellt werden. Trifft dies zu, dann ist die Variable mit der geringeren Auswertesicherheit auszuschließen. Verhältnisskalierte Variablen, die i.d.R. einen nicht linearen Verlauf zeigen (z.B. Baualter) können als Polynom zweiten und dritten Grades in das Regressionsmodell eingefügt werden. Das Verkaufsdatum kann verhältnisskaliert (Dezimaljahr) oder als Jahrgang gruppiert (Empfehlung) untersucht werden.

2.2 Datenanalyse

Liegen ausreichende, vollständige Daten vor und ist das Modell bestimmt, erfolgt die multiple lineare Regressionsanalyse mit einem geeigneten Statistik-Tool. Es werden folgende Werte ausgegeben, wobei hier nur die wichtigsten aufgeführt sind:

Regression

- Das multiple Bestimmtheitsmaß (B), das die Korrelation zwischen Zielgröße und Einflussgrößen misst;
- die Regressionskoeffizienten für alle Einflussgrößen.

Varianzanalyse

Die Rechenergebnisse zur Regression werden durch statistische Test-Größen ergänzt:

- Fisher-F-Test zur Bestimmung der Signifikanz für das multiple bestimmtheitsmaß;
- Student-t-Test zur Bestimmung der Signifikanz der Regressionskoeffizienten.

Residuen

Es erfolgt eine Prüfung der Residuen auf

- Exogenität (der Erwartungswert ist »Null«) und eine optische Prüfung mit geeigneten Graphik-Tools auf
- Homoskedastizität (gleichmäßige Streuung),
- Normalverteilung (Ausreißer können identifiziert, sollten aber nicht automatisch ausgeschlossen werden.).

2.3 Iterationen

Fallen untypische Fälle auf, z.B. Ausreißer bei den Residuen, so sind diese zu überprüfen, ggf. auszuschließen oder im Regressionsmodel neu zu definieren, da sie ein Informationspotential enthalten können.

Nicht signifikante Merkmale können aus dem Modell genommen werden, nicht logische Regressionskoeffizienten (nach Vorzeichen und Größe) sollten sachverständig überprüft werden und können zum Ausschluss von Merkmalen führen.

Im Anschluss sind das korrigierte Datenmaterial und das optimierte Modell, mit weniger oder auch mit neuen Einflussgrößen, erneut einer Regressionsanalyse zu unterwerfen.

3. Typische statistische Kenn-/Prüfgrößen einer Kaufpreisanalyse

- Das multiple Bestimmtheitsmaß (B) liegt zwischen 0 und 1; für Kaufpreisanalysen optimal bei 0,6 bis 0,8.
- Das partielle Bestimmtheitsmaß (partielles R^2) für jede Einflussgröße führt kumuliert zu B und ist auch < 0,2 von Bedeutung.
- Der Korrelationskoeffizient nach Pearson (PK) zwischen Einflussgrößen von < 0,3 ist noch normal (sachverständig prüfen).
- Logische Prüfung der Regressionskoeffizienten nach Größe und Vorzeichen: Plausibel sind z.B. Wohnlage1 > Wohnlage2 oder für Immissionen = JA, dann »Minus«.
- Grenzwerte für t- und F-Test: Signifikanz ist gegeben, wenn < 0,15 (Darstellung in Zahlen oder durch Sternchen, je nach Software).